DOCUMENT RESUME

ED 266 120 SP 027 189

TITLE	Teacher Evaluation Project. The Beginning Teacher Program, Intellectual Skills Development, Validity Studies of the Evaluation System, Special Instrument Development. Report for 1984-1985.
INSTITUTION	Florida Coalition for the Development of a Performance Measurement System, Tallahassee.
PUB DATE	85
NOTE	247p.; For related documents, see SP 027 189-191.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*Beginning Teachers; Classroom Observation Techniques; Data Collection; *Evaluation Criteria; *Program Development; *Research Projects; *Research Utilization; Skill Development
IDENTIFIERS	Florida; *Teacher Evaluation Project FL

ABSTRACT

Reports, summaries, and recommendations are presented on the following research studies: (1) Beginning Teacher Studies; (2) Instructional Skills for Teaching Higher Order Thinking; (3) Development of the Conferential Observation Instrument; (4) Predictive Validity Studies Conducted to Test the Relationship Between Teacher Performance as Measured by the Florida Performance Measurment System (FPMS) and Measures of Student Outcomes; (5) Development of Student Services Personnel Observation System; (6) Development of Data Collection System for Library Media Specialists; (7) Development of Addendum to the Coding Manual for Exceptional Student Education Specialists; (8) Development of the System for Assigning Additional Points in a Master Teacher Program; and (9) A Comparative Study: FPMS with TADS (Teacher Assessment and Development System). Appendices include evaluation instruments used in the studies. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED266120

TEACHER EVALUATION PROJECT

The Beginning Teacher Program
Intellectual Skills Development, Validity Studies of the Evaluation System, Special Instrument Development

Report for 1984-1985

Florida Coalition for the Development of a Performance Measurement System

SP 027189

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B.O. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Conducted under Specific Appropriation 433B, Education Performance Evaluation from the General Revenue Fund to fund Research Development and Evaluation of Effective Performance of Teachers.

Table of Contents

	page
RESEARCH/ADMINISTRATIVE PERSONNEL AND ACKNOWLEDGEMENTS . . .	1
SUMMARIES AND RECOMMENDATIONS.	8
CHAPTER I. BEGINNING TEACHER STUDIES	27
CHAPTER II. INSTRUCTIONAL SKILLS FOR TEACHING HIGHER ORDER THINKING.	56
CHAPTER III. DEVELOPMENT OF THE CONFERENTIAL OBSERVATION INSTRUMENT.	69
CHAPTER IV. PREDICTIVE VALIDITY STUDIES	75
CHAPTER V. DEVELOPMENT OF STUDENT SERVICES PERSONNEL OBSERVATION SYSTEM.	111
CHAPTER VI. DEVELOPMENT OF DATA COLLECTION SYSTEM FOR LIBRARY MEDIA SPECIALISTS.	121
CHAPTER VII. DEVELOPMENT OF ADDENDUM TO THE CODING MANUAL FOR EXCEPTIONAL STUDENT EDUCATION SPECIALISTS . .	131
CHAPTER VIII. DEVELOPMENT OF THE SYSTEM FOR ASSIGNING ADDITIONAL POINTS	135
CHAPTER IX. A COMPARATIVE STUDY: FPMS WITH TADS	143
Bibliography	149

Appendices:

	page
A. BEGINNING TEACHER PROGRAM INTERVIEW	152
B. CONFERENTIAL OBSERVATION INSTRUMENT	158
C. PREDICTIVE VALIDITY STUDIES: RELATIONSHIP OF STUDENT STRATA TO POSTTESTS	159
D. PREDICTIVE VALIDITY: IDENTIFICATION OF OUTLIER IN STUDY B	162

E. PREDICTIVE VALIDITY STUDIES: QUARTILE-BASED SCORING METHOD 165

F. PREDICTIVE VALIDITY STUDIES: SCALING AND INSTRUMENT DEVELOPMENT - DESIGN AND RESULTS. 166

G. STUDENT SERVICES OBSERVATION INSTRUMENT AND TECHNICAL REPORT 173

H. LIBRARY MEDIA DATA COLLECTION INSTRUMENT AND TECHNICAL REPORT 184

I. EXCEPTIONAL STUDENT EDUCATION CONTENT VALIDATION STUDY . . 190

J. CHECKLIST FOR AWARDING ADDITIONAL POINTS FOR TEACHER PERFORMANCE WITH INSTRUCTIONS FOR ITS USE. 199

K. REVIEW OF THE FLORIDA PERFORMANCE MEASUREMENT SYSTEM. . . 207

L. REVIEW OF THE TEACHER ASSESSMENT AND DEVELOPMENT SYSTEM (TADS-MTP) 215

M. COMPARATIVE STUDY OF FPMS/TADS: FREQUENCIES AND DESCRIPTIVE STATISTICS 224

List of Tables

Table page

1. Informed of Legal Requirements. 33
2. Aware of BTP Terms. 34
3. Treatment of Competencies 36
4. Conferences 37
5. Conduct of Observations 38
6. Number of Observations Prior to Post-Observation Conferences 39
7. Who Attended Post-Observation Conferences?. 39
8. Professional Development Plan 40
9. Specificity of Problem Identification 41
10. Rating of Training and Remediation Activities. 42
11. Gain Scores on Summative Instrument 46

12. Summary Outline of 1984-1985 77

13. Correlations Between FPMS Scores and Outcomes. 85

14. Correlations Between Task Engagement and FPMS Scores - Math. 85

15. Correlations Between Task Engagement and FPMS Scores - Social Studies 86

16. Correlations Between FPMS Scores and Posttests - Math . . 87

17. Correlations Between FPMS Scores and Posttests - Social Studies 87

18. Comparison of Higher and Lower Scoring Teachers -- Study A, Achievement 88

19. Comparison of Higher and Lower Scoring Teachers -- Study A, Task Engagement 89

20. Correlations Between FPMS Scores and Task Engagement . . 95

21. Correlations Between FPMS Scores and Achievement. 96

22. Outlier Effects in Correlations 98

23. Post Elementary Teachers Above and Below the 50th Percentile. 99

24. Relationships Between Various FPMS Scalings and Outcomes . 102

25. Combined Z Scores Between FPMS Scores and Task Engagement. 104

26. Combined Z Scores Between FPMS and Achievement 105

27. Combined Z Scores Between Task Engagement and Achievement. 10

28. Elementary Teachers Above and Below the 50th Percentile . 108

29. Post Elementary Teachers Above and Below the 50th Percentile. 108

30. Average Posttest Percentile Ranks. 109

31. Distribution of Respondents by Position, Student Services. 113

32. Distribution of Respondents by Years of Experience . . . 114

33. Distribution of Respondents by School Level 115

34. Opinions of Concept Appropriateness - Counseling. . . . 115

35. Opinions of Concept Appropriateness - Consulting. . . . 116

36. Three Reliability Estimates for Student Services Instrument. 118

37. Distribution of Respondents by Years of Experience, Library Media. 123

38. Distribution of Respondents by Level of Training. 124

39. Distribution of Respondents by School Level 124

40. Distribution of Respondents by Size of School. 125

41. Distribution of Opinions of Concept Appropriateness. . . 126

42. Three Reliability Estimates for Library Media Instrument . 129

43. Distribution of Principals by Districts -- Additional Points 139

44. Distribution of Principals by School Levels 139

45. Distribution of Number of Teachers in School 140

46. Content Validation Data Summary Additional-Point Checklist 141

47. Number of Individuals Falling Above 75th Percentile Cuts . 147

48. Correlations Between EGS and Number of Students in Groups. 160

49. Correlations Between EGS and Number of Students in Strata. 161

50. Plot of Pre and Posttests for Outlier - Study B 163

51. Plot of FPMS Ineffective Scores and Class Posttest Means . 164

52. Relationship of Affect Variables to Student Outcomes . . 170

53. Correlations Among Various FPMS Scalings for Study A . . 172

54. Counseling Scores Across All Services Key Variables. . . 178

55. Consultation Scores Across All Services Key Variables . . 179

56. Scores on Student Services Instrument Across Grade Level . 180

57. Scores on Student Services Instrument Across Session Type. 180

58. Scores on Student Services Instrument Across Group Type . 181

59. Scores on Student Services Instrument Across Counselor Type. 181

60. Scores on Student Services Instrument Across Subject Type. 182

61. Descriptive Statistics for Total and Subscale Scores, Library Media. 187

62. Number of Nonprofessional Staff, Library Media 188

63. Frequency Distribution - Difference Between FPMS and TADS. 225

64. Frequency Distribution - TADS Total Scores. 226

65. Frequency Distribution - FPMS Total Scores. 227

66. Frequency Distributions for Grade Levels 227

67. Frequency Distributions for Elementary and Post Elementary Grades 228

68. Frequency Distributions for Subject Areas 228

RESEARCH INSTITUTE

Principal Investigators

Jean M. Borg
Jeffrey Kromrey
Theodore Micceri
Donovan Peterson
B. Othanel Smith

Staff

Judy Crosby
Brian Gadson
Melva Underbakke
Patricia Zenisek

ADMINISTRATION

Steering Committee

Kathy Bell
Jean M. Borg
Mary Bullerman
Betty V. Fry
John H. Hansen
Laurie Murray
James Parris
Donovan Peterson
B. Othanel Smith
Robert Soar
Raymond Turner
Garfield W. Wilson

ACKNOWLEDGEMENTS

BEGINNING TEACHER STUDY

Coordinators

Mary Jean Davis
John John
Carol Johnson
Ron Kirves
Loretta A. Nielson
Clarence Pilcher
Debby Samec
Roy Scott
Mike Wilkinson
Gladys Williams
Anne Wilson

Interviewers

Jean M. Borg
Mary Bullerman
Arthur Collier
Judy Crosby
James Parris, Director

INTELLECTUAL SKILLS DEVELOPMENT

Investigators

Jean Borg
Mary Bullerman
Judy Crosby
Arthur Lewis
B. Othanel Smith, Director
Melva Underbakke

CONFERRING WITH TEACHERS PILOT STUDIES

West Florida

Larry Blackwell
Barbara Block, Coordinator
Barbara Bruner
Melrose Coleman
JoAnne Gehrdes
Betty B. Hines
Robin Largue
Ronald Peake, Director
Eugene Pettis
Sherman Robinson
Ella Sims
Jo Tarwater
Robert Ann Taylor

West Central

Dakil Alharthi
Arlene Cecilia Chears
Lola B. Dunn
Louis R. Freijo
Leonard E. Gude
Elizabeth A. Harrison
Tawzeya Ismail
Gregory J. Kunka
Donald McClain
David Niederpruen
Donovan Peterson, Director
Brenda V. Sawyer
David J. Steele
Joyce Steinhardt
Armand Ternak

DEVELOPMENT OF VIDEOTAPES

Gayle Adkinson
Celia Anderson
Larry W. Blackwell
Barbara J. Bruner
Jo M. Coker
Melrose Coleman
Jo Ann Gehrdes
Melinda Hendricks
Betty Hines
Robin Largue
Lee Martin
Anna McEwan
Marvetta Nesbitt
Ronald Peake, Director
H. Eugene Pettis
Alison Robinson
Sherman L. Robinson
Ella W. Sims
Jo Tarwater
Robert Ann Taylor
Sammy Taylor
Barbara R. White
Linda Wilson

DEVELOPMENT OF SPECIAL INSTRUMENTS

Student Services Task Force

Juanity T. Borysewich
James M. Eikeland, Chairperson
Jack Foley, Instrument Consultant
John H. Hansen
Jean McGowan
Diane Blum Norton, Researcher
Lane Roosa
Carol Smith

Library Media Data Collection Instrument Task Force

Shirley Aaron
Jack Foley, Instrument Consultant
Tom Hart
Frances Hatfield
Ella Mae Huber
Fred Pfister
Maureen Skinner
Janice Sly, Research Consultant
Jonnie Spriment
Alden Tapio
Sandra Ulm, Chairperson
Melanie Young

Exceptional Student Education Task Force

Betty Carpenter
Sharon Comkowycz
Diane Fardig
Sandra Fradd
Laverne Graves, Chairperson
Catherine Morsink
Noreen Murphy
Doris Nabi
Susan Renick
Jean Rudasill
Roberta Thomas
David Westling

PREDICTIVE VALIDITY STUDIES

Coordinators

Patricia Allen
W. Wade Burley
Miriam Davis
Lynn Gold
Dianne Gordon
Constance Hines
Evelyn Lynn
Linda Murray
Jan Schwartz
Virginia Sharpe
Joyce Swarzman

Observers

Patricia Allen
Lecia B. Altice
Omie Baker
William Beachum
Betty Beiser
Gwendolyn Biddle
Albert L. Bouie
Rowena Brady
Frank Cannella
Flo Davis
Jewel Dickson
Sadie Dretzka
Margaret D. Hartley
Edmund E. Jarvis
Dorothy Johnson
Jean Knight
Gregory L. Lowery
Rosalie Markins
Sam C. Marotta
Kay McLellan
Julie McMaster
Carl Persis
Sherry B. Pigott
Jo Anne Rodkey
Richard Schuler
David Bill Waters
Peggy Wesley
Melanie N. Willette

MASTER TEACHER PROGRAM CONSENSUS STUDY

Coordinators

John Bauer
Dianne Gordon
John Joyce

MASTER TEACHER PROGRAM FPMS/TADS FIELD STUDY

Coordinators

Judy Crosby
Donovan Peterson
Ray Turner
Theodore Micceri
Kimberly DuPree, Clerical Assistant

Observers

Joyce Annunziata
John E. Aycock
Omie Baker
Lenora M. Barr
Harry H. Beil
Betty Beiser
Fred Bertani
Maureen S. Bethel
Mary Brady
William K. Brassard
Frank Cannella
Frazier Cheyney
Barbara Cochran
Fred Damianos
Nelson Diaz
George Fisher
John Gilbert
Dianne Gordon
Sylvia Guilford
Mari-Lynn Hoffman
Steven Ladd
Susan Lebrman
Marion Link
Billye Lowe
James Luideman
Angela Macrina
Jeanine Marcy
Mary Mitchell
Elizabeth A. Murray
Frances K. O'Connell
Douglas Parrish
William Renuart
Judith Richardson
Russell Rowe
Merry Schrage
Jack Silberman
Kenneth Walker
Adelaide Ware
Jean B. Webster
Judith Weiner

CONSULTANTS

Beginning Teacher Program

Mary Bullerman
Dianne Gordon
Carol Johnson
Raymond Jorgenson
Donald McClain

Florida Performance Measurement System

Carolyn E. Evertson
Patricia M. Kay
Joseph Mazur
Donald Medley
Harold E. Mitzel
Robert Soar

Special mention must be made of Patricia Zenisek who did typing and office operations that made it possible to carry out this project.

A project of this magnitude required the involvement of a large number of people in a variety of activities. It was impossible to credit all deserving persons but we have listed those names for which we have a record. We apologize for any omissions. Appreciation is expressed to all of the teachers who participated in the studies described in this report.

SUMMARIES AND RECOMMENDATIONS

Beginning Teacher Studies

The effects of the Beginning Teacher Program have been largely the concern of the State Department of Education and the Steering Committee of the Coalition for the Development of the Florida Performance Measurement System. Two concerns have dominated the discussions of the Committee: 1) operation of the program, and 2) the effects of the program on the performance of beginning teachers and on college programs of teacher education. Program operation is largely in the hands of school districts, whose actions are partly guided by legislation and State Board Rule.

Through the efforts of volunteers from the public schools and universities, the Steering Committee has initiated a program to develop training materials, videotapes, and training activities to prepare principals and other school personnel to use the FPMS Summative Instrument to evaluate beginning-teacher performance.

The Program has been in operation for three years, and three assessments of the Program have been made, as follows:

1. A questionnaire study to learn how the Program is functioning as seen through the eyes of beginning teachers, peer teachers and other supervisors, and principals. (Program Evaluation: Florida Beginning Teacher Program, Department of Education, Office of Teacher Education, Certification and Inservice Staff Development, Tallahassee, Florida, 1983.)
2. An interview study of a sample of 143 beginning teachers in some 10 districts to learn their opinions about the operation of the program. (The findings of this study are reported herein.)

3. A study of the performance of some 1300 beginning teachers, 1983-1984, using the coded behavior at the beginning and end of the school year, to determine the amount and direction of performance modification. (The findings of this study are reported herein.)

These studies are initial inquiries that set the stage for more intensive research on the impact of the Beginning Teacher Program. However, they do provide a rough and ready assessment of the program and suggest important additional research, as well as problems needing immediate attention that can be solved with present knowledge.

Recommendations

1. The orientation procedures should be examined and specific recommendations for improvement suggested.
2. The principle of team support should be examined with the view to recommending changes in its operation or suggesting modification in the principle itself.
3. The practice of professional development planning should be studied and plans for its improvement worked out.
4. The modes of delivering assistance to the beginning teacher need to be assessed and the more effective ones described.
5. The operation of the program should be studied to discover ways of reducing the amount of paperwork for both staff and the beginning teacher.

In addition to the foregoing studies, the Steering Committee also requested that a task force formulate an extended plan of research and development. Accordingly, a group of three individuals was asked to develop a five-year plan for the extension and improvement of the Beginning Teacher Program. The plan was developed and delineated in a flow chart, copies of which are available from the Program Specialists of the Beginning Teacher Program. The research recommended in this Plan is as follows:

1. Research Studies

   a) Studies should be made to set a minimum standard of performance of the essential generic competencies.

   b) A study of the portfolio system should be made, including content and the relationship between the portfolio and the beginning teacher's success as measured by summative observations.

   c) A study of the effectiveness of inservice activities and materials used in BTP should be conducted.

   d) Study of the impact of BTP upon the performance of the beginning teacher should be continued.

   e) A study of the use of funds now allocated to the BTP should be undertaken.

   f) The state legislature has enacted legislation establishing an alternative teacher preparation program that enables a four year graduate of an accredited college to teach while undergoing training by a school district. In view of this alternative, it is appropriate to ask whether or not such a

BEST COPY AVAILABLE

program of preparation produces competent teachers. To answer this important question, research should be undertaken to ascertain (1) the type and extent of preparation provided by the districts and how it compares with the BTP; (2) how the performance of the alternative teachers compares with that of the beginning teachers as measured by the summative instruments of the FPMS; and (3) the effect of the alternative program as determined by the criteria specified in the legislation.

2. Policy Studies

a) Assuming that observations of classroom performance should be made by trained observers, who should these observers be and by whom should they be trained?

b) By what criteria should observers be certified and for what period should the certificate be valid?

c) How many summative observations should be required for each beginner and by how many observers?

d) Should legislation or State Board Rule be sought that would enable the Department of Education to require submission of all observations made with summative instruments?

e) What should be the role of associate master and master teachers in servicing the BTP?

f) How many formative observations should be made of each beginning teacher? Should the teacher's competence be judged by performance in each domain or by summative observations alone?

g) What preparation should be provided for the support staff to enable it to function effectively?

h) What should be the role of college personnel in the BTP?

Instructional Skills for Teaching Higher Order Thinking

There are at least two components of the content of instruction: 1) concepts, laws, rules, values, and facts; and 2) intellectual skills and the procedures and techniques of manipulating the subject matter in order to identify and demonstrate relationships among the components of instructional content. The first of these is explicitly recognized in the objectives of instruction and in measured outcomes of instruction. The second is taught incidentally, if at all, and with little or no recognition of these skills in the evaluation of student achievement.

The reasons for the neglect of these intellectual skills and procedures are many, but they boil down to the fact that little or no instruction in them is provided either in programs of teacher education or in the liberal arts curriculum. Furthermore, the research on the pedagogy of these skills and procedures, while extensive, lies fallow in research journals. The results of this research can be used by teachers only if it is compiled, organized, and entered into a form suitable to the requirements of classroom instruction.

A task force of the Teacher Evaluation Project is undertaking the foregoing task. This group has compiled a bibliography of some 400 studies and begun the assimilation of the findings into categories of types of problems, as follows:

1. Partly-structured problems, identified as perplexities where known concepts, principles, and procedures are not sufficient to generate a solution and where criteria for deciding among probable solutions are debatable.
2. Classification problems, identified as situations that require formulation and testing of criteria for grouping objects, events, etc.
3. Evaluation problems, identified as situations in which one person is required to evaluate what another expresses as a conclusion, or sets forth as an argument.
4. Well-structured problems, identified as problems for which there are agreed-upon rules, formulas, and procedures for solving them and criteria for telling when they are solved.

The task force has developed sets of concepts and indicators of effective teaching practices and assembled the research for a considerable number of these concepts and indicators.

Recommendations

To complete the task of assembling the research and refining the concepts and indicators, and to develop training materials and instruments for classroom observation to measure teacher performance in the teaching of intellectual skills and procedures, it is recommended that provisions be made to:

1. Complete compilation of the research findings.

2. Continue work on the extension and refinement of the concepts and indicators of effective teaching practices.
3. Develop formative and summative observation instruments.
4. Develop a manual for teachers, setting forth the essential knowledge about intellectual skills.

Development of the Conferential Observation Instrument

The Beginning Teacher Program is partly based on the proposition that peer teachers, and other appropriate school personnel, have the skills to establish rapport with beginning teachers and assist them in positive and constructive ways with classroom problems. However, the skills to gain confidence of others and to establish and maintain positive working relationships must be learned. Many teachers and other school personnel have not acquired the skills of effective personal interaction. Research and development work was done this year to provide the knowledge base and instrumentation needed to prepare teachers and others to assist beginning teachers.

The first step of this project was to locate the research on conferring and consultation. While quantitative research is limited, it yielded enough solid information to identify basic concepts and indicators of a considerable number of skills. The concepts and skills were consolidated and the resulting instrument was content-validated, and has undergone preliminary testing for reliability.

As a result of this year's work, the following products are available to school districts and Colleges of Education:

1. An instrument on conferring.
2. A coding manual for the instrument.
3. A set of videotapes for training personnel in the use of the instrument.

Recommendations

In order to implement the program for the preparation of peer teachers and others in the skills of conferring, the following recommendations should be considered as basic elements:

1. Development and testing of training programs in the conferential skills that would be available to master trainers so that they can help districts train their personnel in conferring.
2. Development of additional training videotapes for practice during training.
3. Development of procedures for certifying peer teachers and administrators in conferential skills. This requires: a) development of cognitive examinations on the knowledge base, and b) development of criteria videotapes for testing observer accuracy.

Predictive Validity Studies

Developmental work on the Florida Performance Measurement System (FPMS) is in its fifth year. Previous studies established content

validity of the domain documents, instrument and observer reliability and generalizability across frame factors and instruction formats. The FPMS was normed on a representative sample of Florida's teachers. Exploratory predictive validity studies were conducted this year to estimate the relationship between FPMS scores and measures of student engagement and achievement. These studies also examined the appropriateness of quartile-based scoring and alternative scoring procedures for this instrument. Five studies were started, two of which were completed by the time this report was written. Of these two studies, a two-week study (N = 1100 students, 51 classes) was conducted in elementary reading and mathematics and a one-semester study (N = 1050 students, 40 classes) was conducted in secondary social studies.

Stratified class means were used as the unit of analysis (high, average and low achievers). Task engagement data were collected by observers. Content-specific tests were used as pre and post measures of achievement in the elementary study. A general achievement test was used as a covariate in the secondary study, with a content-specific test as a posttest. Meta-analytic techniques were used in analysis. Summary findings of predictive validity studies to date suggest the following:

1. There is a highly significant relationship between FPMS scores and task engagement;
2. There exists a highly significant relationship between task engagement and student achievement;
3. There exists a significant relationship between FPMS scores and student achievement at the elementary level; and

4. Teachers scoring at or above the FPMS norming group's 50th percentile probably have more positive effects on their students' achievement than those scoring below this point. (This effect is established at the .06 level of confidence.)
5. Quartile-based scaling, the scaling used for the Master Teacher ranking, is probably the most stable estimate of the instrument's relationship with student outcomes.

Analyses to date suggest that the more low achieving students in a class, the worse the low achieving students do; and the more high achieving students in a class, the better low achieving students do.

Recommendations

1. Results support the use of FPMS in the Beginning and Master Teacher Programs.
2. Studies conducted this year were exploratory and the findings are tentative. Further study should be conducted to investigate the predictive validity of the FPMS.
3. The following factors should be considered when planning future predictive validity studies of the FPMS (Kay, et.al. Review Report, 1985).
   a) Increase the number of observations to insure a representative sample of each teacher's behavior.
   b) Conduct additional long-term studies (one-semester minimum) to better establish teacher effects on student outcomes.
   c) Improve procedures for collecting student task engagement data.

d) Identify additional subject-specific, district-wide standardized tests for use as measures of achievement.

Development of Student Services Personnel Observation System

It was recognized at an early date that FPMS, based on research on effective classroom practice, was inadequate for observing and evaluating student services personnel: occupational placement specialists, guidance counselors, school psychologists, social workers, and student services administrators. Although classified as instructional personnel, these persons do not typically work in instructional settings, but rather in consultative and diagnostic settings.

Following the same steps as for the development of the other components of FPMS, a task force of student services leaders compiled a set of six concepts and their performance indicators for counseling, and eight concepts with their indicators for consultation. Content validity was established through literature search, judgment of 464 individuals active in the field, and the consensus of the review team.

A single observation instrument with 19 effective and 18 ineffective indicators sufficed for both counseling and consulting areas. Reliability studies used the generalizability approach, with an additional determination of internal consistency. Reliability coefficients were calculated for effective indicators, ineffective indicators, and total score. Only total scores should be used for these

instruments since reduced variance and low reliability estimates were found with ineffective indicators because of their relatively rare occurrence. For total score, intraclass discriminant reliabiltiy was r = .58; coefficient alpha (internal consistency), r = .78; and intercoder agreement, r = .56.

Available materials for training observers in the student services field include domain documents seven and eight of the FPMS, a coding manual, observation instrument, and videotapes for training and for criterion assessment.

Recommendations

1. Concepts and indicators based solely on consensus should be tested empirically for their effects on client populations.
2. Further studies of reliability, generalizability, and practicality should be conducted, using data from the Master Teacher Program.

Development of Data Collection System for Library Media Specialists

The concern of the Task Force for Library Media Specialists was that while the FPMS summative instrument is appropriate for assessing classroom performance of media specialists, it does not address their non-instructional functions. Using both professional consensus and review of literature, a group of library media specialists identified nine concepts and thirty-two indicators, with examples and supportive

BEST COPY AVAILABLE

evidence, which were organized into Domain 9 of FPMS. A cross-section of 727 library media specialists representing all levels, sizes, and organizations of schools and all levels of experience responded to the content validation study with a very high level of agreement on the items included.

The nine concepts were classified under three functions -- management, collection-development, and services -- and a data gathering instrument was designed. Only positive indicators appear on the instrument, since the behaviors recorded either do or do not occur. Hence, absence of tallies indicates absence of the behavior.

Reliability studies indicated that the total instrument score was the most reliable. (Total instrument: $r = .77$ to $r = .89$; management section: $r = .71$ to $r = .81$; collection-development and services sections: $r = .63$ to $r = .70$.) Therefore, total score should be used for evaluation. Products developed for training include the domain document (Domain 9 of the FPMS), the coding manual, and the data collection instrument.

Recommendations

1. Concepts and/or indicators based on consensus alone should be empirically tested.
2. Further testing of reliability, generalizability, and practicality should be conducted, using the Master Teacher Program data.

Development of Addendum to the Coding Manual for Exceptional Student Education Specialists

A Task Force examining the appropriateness of the Florida Generic Competencies and the FPMS indicators for teachers of exceptional students found, through practitioner consensus and review of research, that the special competencies and indicators they identified were in fact special cases of the generic concepts and indicators in FPMS. (They had been so classified in the original Beginning Teacher Competency-Indicator match.) As a result of these findings, the Task Force agreed that the only modification of the system for exceptional education teachers was a rewriting of some of the indicators in the Coding Manual to ensure their applicability for observing teachers who are working with exceptional students. This was done.

The resulting coding manual was submitted to 225 practitioners for content validation, and data are being monitored to determine effects on reliability and norming distributions.

Recommendations

1. Master Teacher Program data should be used to verify reliability and generalizability of FPMS in the specialized areas of exceptional student education.
2. A study should be conducted to determine whether additional indicators represent an unfair advantage for ESE teachers over regular teachers in Master Teacher Program evaluations.

Master Teacher Program: System for Assigning Additional Points

The legislature provided for additional points to be added to performance evaluations in the event the principal determines the evaluation system does not adequately recognize applicant performance. To partially standardize this procedure (a basic requirement of state-wide programs of evaluation), an instrument was developed containing items from research, district evaluation instruments, and the Florida competencies. The instrument was content validated by a sample of principals (n=106).

The scoring procedure was left substantially to the discretion of principals; up to ten points could be assigned to a candidate for one item or spread among several items. Reasons for assigning points had to be documented by principals and approved by the Department of Education. By Board Rule, the ten points, or any proportion thereof, were scaled to each instrument used in the MTP, allowing a maximum of 10 percent of the median score. This procedure, however, did not consider variations in instrument score distributions. In application, just as a result of statistical artifact, it could make substantially more candidates eligible under one system than another.

Recommendations

1. Attaining equality of evaluation among candidates, a basic tenent of teacher evaluation, is compromised when principals can assign ten points to one category while others spread these points among a larger number of categories -- up to fifteen. This procedure

should be carefully examined and alternative procedures tested to better standardize the value of these points across candidates.

2. Distributions of performance scores on each instrument (i.e. FPMS, TADS, Library Media and Student Services) from the 1985 Master Teacher Program observations should be examined to determine the power and effects of additional merit points. THIS IS A CRITICAL RECOMMENDATION since variations in shapes of distributions can cause the same number of additional points to qualify a substantially larger proportion of candidates under one system than under the other.

A Comparative Study: FPMS with TADS

Since two systems are used to evaluate teachers in the Master Teacher Program, TADS in Dade County and FPMS in all other counties, results of their application must be compared to determine 1) whether these instruments assess similar dimensions of teacher behavior, 2) whether they rank order teachers similarly on a scale of effectiveness and 3) how they relate to measures of student outcome.

Plans to address the first of these three areas of inquiry were made this year, and a field study was conducted to address the second area. This study compared FPMS coding with TADS rating, and examined their respective rank ordering of teachers on a scale of effectiveness. It compared 66 teachers in 12 grade levels. The teachers were observed at the same time using TADS and FPMS observation instruments. The single correlation between scores for the two instruments was moderately

significant (r = .49). This relationship suggests that in a research context, TADS and FPMS assess similar dimensions of teacher behavior. Overlap at the 75th percentile was approximately 55 to 57 percent; that is, about 55 to 57 percent of those individuals scoring at or above the 75th percentile on TADS also reach that point on FPMS, and vice versa.

Cross-validation of FPMS/TADS items is scheduled to be conducted in early fall of this year.

Recommendation

1. Continue research on FPMS/TADS according to plans developed this year.

TEACHER EVALUATION PROJECT REPORT

Chapter I

Florida Beginning Teacher Studies

A graduate of a well-known university begins her first day of teaching. She meets her first class, a group of thirty students in biology; a group of strangers, some of whom have no self-control and about whom she knows nothing save what the impersonal school records tell her. While thinking about what she plans to do, everything seems to happen at once. Some students are scrambling for desks, others are jabbering, and a few sit quietly at their chosen desks. For a moment, she ponders the situation. She has planned the lesson and now she wonders how to begin, what to do in order to gain and retain the class's attention. At the same time, the gnawing thought that this is the first of today's five classes flickers through her mind.

To face a group of thirty strangers, to know that you are responsible for their immediate conduct and their personal and intellectual development is an experience the like of which few individuals ever undergo. It can be shattering or personally rewarding, depending upon the course of events over which the individual, especially a beginning teacher, must take control to be successful.

Ms. Boggs had been an honor student and anxious to become a teacher. She had majored in science with a minor in mathematics. In addition, she had completed the specified program in pedagogy. Her performance as a student teacher was well above average. Yet her first day as a full-time teacher filled her with anxiety.

As days turn to weeks and weeks to months, Ms. Boggs does not find the excitement of teaching she anticipated. In her more difficult hours, she turned to her colleagues only to learn that, being occupied with their own work, they had little time to assist her. As for the principal, he was not easy to approach and what she learned from him was of little help. As the school year wore on, Ms. Boggs daily became more and more concerned about her choice of teaching as a career.

Why should this individual who has prepared herself to teach find her first day of teaching upsetting? And why should the first day's fear continue through the year, softened only by the development of a few survival skills and the alleviation of tensions that comes from familiarity with the system and its people? Is it because of inadequate preparation? Is it because a year of on-the-job experience before plunging into full-time work, as in other learned professions, is needed to round out one's preparation? Is it because organizational expectations and other aspects of the work environment are almost overwhelming to the uninitiated? All of these conditions, and more too, evoke the anxiety suffered by beginners in the teaching profession.

Beginning Teacher Program: Inception and Purpose

To alleviate these conditions, the Florida legislature established the Beginning Teacher Program in 1981, effective July 1, 1982. The legislation enacting this program required successful completion of a year-long program as a condition of certification. The anticipated results of the program were 1) more effective classroom performance, 2)

increased student achievement, and 3) decreased beginning teacher attrition.

It is three years since the Beginning Teacher Program became operational. In these years, the number of beginning teachers has constantly increased. Today there are approximately 7,000 beginning teachers, including a number with varying amounts of experience. It is appropriate to ask: How fare beginning teachers? Have they received the support required by legislation and State Board Rule? What impact has the Beginning Teacher Program had upon their classroom behavior? Has the drop-out rate of beginning teachers decreased since the program's inception? Only provisional answers can presently be given to the first two questions and no information is currently available on the beginner drop-out rate.

Study A - Interview Study of District Programs

What information and what assistance is the beginning teacher receiving? In order to answer this question, an interview study of beginning teachers was conducted during April 1985.

The interviews were conducted by five trained interviewers. To facilitate uniformity of procedure and data categories, an interview instrument was developed. It consists of some 100 items classified into the following groups: demographic information, orientation, and implementation. For purposes of data analysis, these categories were subsequently broken down into the following elements:

1. Legal requirements,
2. Personal involvement,
3. Promptness of implementation,
4. Observation and preparation of professional development plans, and
5. Modes of training.

The interview instrument was revised by the interview team and copies provided to each member of the team after the interview procedure was formulated (Appendix A).

Sample

The sample included 143 beginning teachers in 10 school districts consisting of four Panhandle districts and six central Florida districts. Two of the ten were large urban districts, three intermediate, and five rural or semirural districts. The number of teachers sampled in each district ranged from 11 to 15. All grade levels were represented -- 57 elementary, 34 middle school, and 51 secondary teachers.

Eighty teachers, or approximately 56 percent, received their teacher training in Florida institutions, and 73 of these, or 91 percent, were trained in Florida's major state universities. Sixty-three of the 143 teachers, or approximately 44 percent, were trained in non-Florida schools.

One hundred twenty-six, or 88 percent of the teachers, had no prior teaching experience. The remaining 17 teachers, or approximately 12 percent, reported from one to 22 years of teaching experience. Seventy percent of the interviewees were female, and 86 percent were white non-Hispanic.

Procedure

Interviews of beginning teachers covered four areas: identifying data, the beginning teacher's understanding of the Beginning Teacher Program, the procedures that were followed, and the type of help the beginning teacher received. The interview format included questions to guide the information gathering through these four areas, and it provided room for additional comments.

Instructions to interviewers included the following:

1. Briefly explain your role as data collector for the evaluation of the Beginning Teacher Program. Inform the beginning teacher of his/her role in the interview, emphasizing that all compiled data will be anonymous and confidential. Give them a specific statement like the following to start off the interview: "I wish to emphasize that all information gained from this interview will be anonymous and confidential and only considered as part of grouped data. No individuals will be singled out. You were selected as part of a random sample of beginning teachers for this interview."

2. Give the teacher an overview of what areas will be covered and indicate your willingness to add information, particularly in the summary, so they know everything will be covered and are less likely to digress.
3. Where appropriate, circle numbers that represent the beginning teacher's responses.

Presentation of the interview data is organized under two headings: orientation of the teachers to the program and program implementation and operation.

Orientation to the Program

Local school systems usually provide information about the school, its policies, and regulations. Such information is updated from year to year for experienced teachers, and is given to new teachers, both experienced and inexperienced, to facilitate their understanding of the system. While information about the school system is conducive to a properly functioning school, the Beginning Teacher Program generates additional needs for information such as sources of assistance, legal requirements, and the meaning of key terms (for example, portfolio, competencies, and performance evaluation).

The legal requirements of the program should be known by every beginning teacher, for without this information the teacher does not know his or her legal rights and obligations. It can be seen from Table 1 that more than 97 percent of our sample was informed of the legal requirements, such as appeal process and essential competencies; that a

BEST COPY AVAILABLE

little over 42 percent gained their information from the building level administration (BLA), and that 51 percent were informed by other sources.

Most of the beginners, 79 percent, acquired the legal information through workshops, lectures, or meetings; some four percent got their information by informal conversation and other means. The number receiving information by handouts may exceed the number reported in Table 1, for handouts about important requirements are typically provided in workshops.

TABLE 1

Informed of Legal Requirements

	Frequency	Percent of 143 Cases
Informed		
No	3	2
Yes	139	97
No Answer	1	0
By Whom Informed		
BLA	60	42
Peer Teacher	5	4
Specialist	1	0
Others: workshop personnel, program directors or coordinators,etc.	73	51
No Answer	4	3
How Informed		
Workshop/Lecture/Meeting	113	79
Handout	20	14
Informal Conversation	4	3
Other	2	1
No Answer	4	3

Part of the orientation of a beginning teacher is ensuring familiarity with the terms used in the program. It is clear from Table 2 that the teachers represented in this sample were for the most part made aware of commonly used terms of the Beginning Teacher Program. The expressions with which they were least familiar are "Florida Performance Measurement System" and "Professional Development Plan". Twenty-nine percent reported that they were not made aware of the first expression, and 18 percent said they were unfamiliar with the second. Other terms such as portfolio, summative instrument, formative instrument, support staff, and peer teacher were familiar to over 94 percent of the interviewees.

TABLE 2

Aware of BTP Terms

	Frequency	Percent of 143 Cases
Made aware of term PDP?	117	82
Made aware of term FPMS?	101	71
Made aware of term Portfolio?	134	94
Made aware of term Summative Instrument?	136	95
Made aware of term Formative Instrument?	134	94
Made aware of term Support Staff?	139	97
Made aware of term Peer Teacher?	141	99
No Answer	1	00

A central feature of the Beginning Teacher Program is the demonstration of Florida Minimum Essential Competencies, which is the primary requirement for State Certification. It is therefore important

that all teachers be thoroughly informed about these competencies and that they understand them and become proficient in their use.

To what extent are these competencies made available and explained during the orientation period? As shown in Table 3, 94 percent of the beginners were made aware of these competencies. This means that six percent or some seven teachers of the study sample were not made aware of this essential feature of the program. The competencies were made available in written form to 93 percent of the teachers, but were reviewed and explained for only 78 percent. Apparently almost 25 percent of the teachers were given no instruction in this fundamental component of the program. This is a deplorable deficiency, indicating a lack of concern about instructional competence, which the profession should overcome at all cost.

Table 3 shows that only 32 percent of the beginning teachers had completed the program when the survey was made in late April. This suggests that almost 70 percent were still receiving the assistance of the support team, a fact that lends support to the need for a year-long program.

TABLE 3

Treatment of Competencies

	Frequency	Percent of 143 Cases
Made aware of Florida minimum competencies?	135	94
Competencies reviewed and explained?	112	78
Competencies available in writing?	133	93
Were competencies identified before observation?	131	92
Signed off on competencies?	45	32

Operation of the Program

This phase of the program emphasizes observation, planning for professional development, and remedial activities.

The observational component consists of pre-observation conferences, performance observation, and post-observation conferences. The purpose of the pre-observation session is to prepare for actual observation by further discussion of the competencies and by arranging a convenient time for the observation. Extracting from Table 4, over six percent of the teachers reported that no such conferences were held. Almost all teachers (92 percent) were informed of competencies prior to observation, consulted about appropriate times for observation (97 percent), and told when observations were to be conducted (96 percent).

Table 5 shows that 90 percent of the first screening observations were made by building level administrators, and that peer teachers and district personnel conducted almost all of the remaining first observations and in about equal proportion.

TABLE 4

Conferences

	Frequency	Percent of 143 Cases
Pre-observation conference	134	94
Competencies introduced before observation	131	92
Consulted about observation	138	97
Advance notice of observation	137	96
Post-observation conference held	142	99

The first screening observations were made at various times throughout the first semester, as is evident from inspection of Table 5. Thirty-nine percent were made during the first month of the school term, 37 percent during the fifth and sixth weeks, 18 percent during weeks seven through nine, and six percent after more than nine weeks. It is important to note that about 25 percent of these beginning teachers were delayed in identifying and correcting whatever skills needed improvement until almost one fourth of the school year had elapsed.

The frequency of observations is impressive, exceeding in most areas the state requirement. Almost 83 percent of the teachers, as is evident from analysis of Table 5, were observed five times or more. Nine observations were made of approximately one teacher in four.

Referring back to Table 4, post-observation conferences were held in 99 percent of the cases, although 30 percent were not held until the second observation, as is evident from Table 6, thus possibly delaying

TABLE 5

Conduct of Observations

	Frequency	Percent of 143 Cases
Who made first screening observation?		
BLA	128	90
Peer Teacher	6	4
Specialist	1	0
County Personnel	8	6
First screening observation made		
1st 4 weeks	55	39
5-6 weeks	53	37
7-9 weeks	25	18
9+ weeks	9	6
No answers	1	0
Number of times observed		
1	1	0
2	3	2
3	7	5
4	12	9
5	31	22
6	18	13
7	16	11
8	19	13
9	34	24
No Answers	2	1

the planning of professional development programs. However, if these conferences were held after the second observation in order to assemble more adequate data for planning, the delay would seem to be justified. Three percent of the post-observation conferences were attended only by peer teachers, with about 30 percent only by building level administrators, and about 23 percent only by other support staff, as shown in Table 7.

TABLE 6

Number of Observations Prior to Post-Observation Conferences

Conference	Frequency	Percent of 143 Cases
1	62	44
2	43	30
3	4	3
4	33	23

TABLE 7

Who Attended Post-Observation Conference?

	Frequency	Percent of 143 Cases
All Staff	62	43
BLA	43	30
Peer	4	3
Other	33	23
No Answer	1	1

All can agree that a tentative professional development plan, based upon observation data, should be sketched during the first post-observation conference. It is reasonable to suppose that the plan is more likely to be successful when developed in cooperation with the support staff. It is interesting to note that only a little more than half of the development plans were cooperatively worked out, as shown in Table 8, although the teachers had input into their plans 69 percent of the time. Thirty percent of the teachers reported that plans were worked out without the use of observation data. Nevertheless, the

teacher was informed of the problems indicated by the data 76 percent of the time. In the course of the year, the plan was revised for 64 percent of the teachers.

TABLE 8

Professional Development Plan

	Frequency	Percent of 143 Cases
PDP written as team?	80	56
Observation data used in PDP?	99	69
Knew problems to solve?	109	76
Input in PDP development?	99	69
Was PDP revised?	92	64

The Training Phase

As shown in Table 9, training needs of more than half of the teachers were identified at the indicator level. Needs were identified at either the domain or concept level about 27 percent of the time. That 19 teachers gave no answer calls for an explanation. Could it be attributed to the absence of problems, to inability to recall, or to some other factor?

The interview format called for the teacher's perception of the relative value of seven training activities as follows: focused workshop, general workshop, consultation/demonstration, guided observation of skilled teachers, independent work, specific instruction by team members, and professional reading. The teachers were asked to

TABLE 9

Specificity of Problem Identification

	Frequency	Percent of 143 Cases
Domain	20	14
Concepts	19	13
Indicators	77	54
No Answers	27	19

rank order the three most useful activities. These ranks were then combined across teachers to arrive at the number of times each activity was chosen, and the mean rank. Ratings were developed from the frequency and level of ranking, as shown in Table 10. These were computed by dividing the absolute number of rankings by the mean rank. Thus an activity that frequently received a high rank (1 is high, 3 is low), received a high rating (i.e. specific instruction = 73/1.73 = 42.2), and one receiving few low ranks got a low rating (i.e. independent work = 31/2.52 = 12.3).

Table 10 depicts the ranking of these activities. It is clear that specific instruction is by far the most preferred, with consultation and observation of skilled teachers also receiving some support. Independent work, professional reading and workshops received little support from the beginning teachers.

TABLE 10

Rating of Training and Remediation Activities

Activity	Rating	Number of Times Ranked	Mean Rank
Specific Instruction	42.2	73	1.73
Observing Skilled Teacher	28.7	52	1.81
Consultation	26.8	51	1.90
Workshops, Focused	18.8	35	1.86
Professional Reading	17.0	35	2.06
Workshops, General	14.7	29	1.97
Independent Work	12.3	31	2.52

Interviewer Input

The interviewers recorded information about the Beginning Teacher Program expressed by the beginners themselves. This information constitutes the substance of the foregoing pages. Over and beyond this information are certain insights of the interviewers that occur from nuances, hints, side remarks, and outright claims of the teachers. While the reliability of this information is somewhat fragile, it adds a tincture of reality that lends significance to the interviewee data. These insights are summarized in the following paragraphs.

1. Generally speaking, beginning teachers think the program is beneficial, that the purpose of the program is good, but that certain parts, such as documentation of competencies or keeping a portfolio, should be better implemented or changed.
2. Beginning teachers definitely sense attitudes of the administration toward the program. Where the attitude is

negative, the beginner expects and gets little or no assistance. Some beginners claimed that no one did anything. They were simply told to fill out required forms, and, after doing so, were signed off as okay.

3. The practice of team assistance is apparently not recognized by many beginners, a fact that causes one to wonder about the extent to which the principle of team assistance is followed or whether the principle is workable in the school setting.

4. Many beginning teachers appear to know little about the professional development plan. Some say they never heard of it. Others know about such a plan, but quickly say that basically the plan was worked out for them with little chance for input, even when the opportunity to do so was offered. Some considered their portfolio to be their professional development plan.

5. The amount of paperwork required of the beginner is a cause of much complaint. For example, many said that they spend an inordinate amount of time trying to match notes and other products with the 24 competencies. Then they wonder what becomes of their documentation, for no feedback is given. A few beginners wondered whether some of the paperwork was really their responsibility. At any rate, it was generally felt that goals of the program were threatened by extensive paperwork.

6. Orientation to the program at the beginning of the year is considered desirable by most beginners, and some appear to prefer small doses of information and workshops over time rather than in toto at one or two sessions.

Summary

This study is designed to obtain the reactions of beginning teachers to the assistance they receive from the schools and districts in which they teach. The Beginning Teacher Program has been in operation for three years, ample time for most of the district and school level difficulties to have been worked out.

The findings of this study were derived from interviews with a sample of 143 beginning teachers selected from 10 school districts representing the panhandle and central Florida. The interviews were conducted by five trained interviewers, using an inventory to limit variation in the content and procedures of the exchange between teachers and interviewees.

Generally speaking, the beginners were given information about the legal requirements they must satisfy and the various elements of the Beginning Teacher Program. However, certain crucial information either did not reach the teachers, or else, the explanations given them were inadequately presented. For example, 93 percent of the teachers received the compentencies in written form, although 25 percent reported that they were given no instruction in the competencies. Another fact of crucial importance is that over six percent of the teachers were not involved in post-observation conferences, leaving them to wonder about the results.

Moreover, the team concept whereby beginners are given benefits of peer teachers, principals, and supervisors is apparently not being

implemented in many cases. Only three percent of the cases had the advantage of peer teacher assistance in the post-observation conference. Furthermore, one wonders how peer teachers can be of much assistance when they have no opportunity to assist in identifying the beginners' needs at the outset. The fact that less than half of professional development plans were cooperatively worked out reinforces one's doubt that the team approach is being utilized.

Modes of delivering assistance to beginners is a subject about which there are many opinions. It is interesting to note that our sample reports that the most helpful modes of giving assistance are specific instruction, observing skilled teachers, and consultation in that order.

Study B - Effects of Program on Teacher Performance

It is one thing to know what assistance the beginning teacher receives; it is quite a different thing to know what changes occur in the teacher's behavior during the first year's experience. The study here reported is a preliminary exploration of these changes in behaviors.

Data for this study were drawn from the 1983-1984 school year. We can give only provisional answers to the foregoing questions since adequate data are not available, due in part to the program's loosely coupled components. For example, we do not know the length of observation periods, whether any two observations of a teacher were made by trained observers, nor the frame factors that condition the scores.

Furthermore, it is possible that in some cases the beginners have been trained in FPMS prior to observation. For these and other reasons we must view the findings of our study of the 1983-1984 beginners more critically than we would otherwise do.

In 1983-1984, thirty-five districts contributed pre- and post-observation data on 1002 teachers. Analyses of these data indicate that significant gains were registered for both sides of the summative instrument, as shown in Table 11. District by district analysis indicated consistent positive results on the ineffective scale, with 28 of the 35 districts (80 percent) showing gains. On the effective scale, 23 of 35 (66 percent) showed positive gains.

TABLE 11

Gain Scores on Summative Instrument

	T	dF	P
Effective Indicators	4.64	1000	.0001
Ineffective Indicators	11.17	1000	.0001

Analysis of gains yielded significant differences across district size. Larger districts showed greater gains than smaller districts, and more affluent districts showed greater (although not significantly so) gains than less affluent districts.

It is of some interest to note that in an opinionnaire study, where it was possible to combine 47 cases of the survey study with summative

observation scores, a significant positive relationship was found between beginners' perceptions of preservice internship quality and summative scores for Domain 3, "Instructional Organization and Development" ($r = .33$, $N = 46$). Furthermore, beginners' perceptions of preservice preparation in management of student conduct correlated positively with summative scores on Domain 2, "Management of Student Conduct" ($r = .30$, $N = 47$).

This study was designed to ascertain, among other things, the comparative effects of the various Colleges of Education upon the initial performance of beginners as measured by the summative instrument of FPMS. This aspect of the study proved to be inappropriate, for the available data did not permit us to determine that the observations were made by trained observers. In addition, observations were made by only one individual, thereby impairing the reliability of the observations. In future studies it is hoped that these defects can be remedied, and that comparative studies of the influence of college programs can be conducted.

Recommendations

It is evident from both interview data and the supplementary information of the interviewers that, generally speaking, the program is being implemented, although there is need for improvement in several fundamental respects. In order to focus attention on these needs and to suggest ways to meet them, the following recommendations are made.

1. The orientation procedures should be examined and specific recommendations for improvements suggested.

Justification: There is large variation in the personnel responsible for conducting the orientation. This is highlighted by the fact that orientation was given by building level administration in 43 percent of the cases and that in almost an equal number of cases the teachers could identify those who informed them about the program by only the term "other", meaning perhaps district personnel.

While the essential competencies were explained to a large proportion of the teachers, 22 percent were not given this advantage; and a considerable number of beginners had little information about the professional development plan and 18 percent said they had never heard of the expression.

2. The principle of team support should be examined with the view to recommending changes in its operation, or suggesting modifications in the principle itself.

Justification: There appears to be little interaction between the building-level administration and the peer teacher regarding the initial screening observation. The peer teacher is seldom involved in the post-screening observation conference where an analysis of the observation data supposedly occurs. Furthermore, there is little evidence that supervisors or other personnel are involved.

3. The practice of professional development planning should be studied and plans for its improvement worked out.

Justification: From interviewer data and supplementary information, it is clear that the principle of professional development planning is not being adequately carried into effect. It seems clear that the whole notion needs refinement with regards to purpose, who is responsible, who should participate, and who should be responsible for supervision of its implementation.

4. The modes of delivering assistance to the beginning teacher need to be assessed and the more effective ones described.

Justification: This is a significant component of the program, for the effects of the program as a whole depend upon the success of the training activities. The beginning teacher ratings of various modes of providing assistance suggest that in their experience certain types of activities are much to be preferred over others.

5. The operation of the program should be studied to discover ways of reducing the amount of paperwork for both staff and the beginning teacher.

Justification: Although the interview schedule did not provide for this item in the interviews, excessive paperwork appeared to be a general complaint. Simplicity would appear to be justified at least for the beginner who is already burdened

with the task of adjusting to the demands of full-time responsibility.

Beginning Teacher Program: A Developmental Plan 1985-1990

From time to time, the Beginning Teacher Program (BTP) has been the subject of much concern to the Coalition's Steering Committee. Among the more persistent concerns have been the lack of training materials, lack of training in the use of FPMS formative observation instruments and inadequate preparation of peer teachers and others in the procedures and techniques of delivering assistance to beginners. These concerns culminated in a request that a long-range plan for the further development and evaluation of BTP be worked out and presented to the Steering Committee. As a response, a five year plan was developed cooperatively with the DOE Program Specialists of BTP, and is presented in the following pages. The recommendations provided in this plan are compatible with and reinforced by the studies reported in the foregoing sections.

The Florida Beginning Teacher Program is a continuation of preservice preparation in the form of a modified internship. While it includes experienced teachers, newly employed, it does so to verify their competence and to familiarize them with the Florida program.

The effects of the BTP depend upon the articulation and mutual reinforcement of three components: school districts, Colleges or Departments of Education, and the State Department of Education. The

professional preparation is initiated by the colleges, continued by school districts in the first year of employment, and further extended and enriched by the colleges at the post-baccalaureate level and by the districts through inservice activities. The State Department of Education, chiefly through the Office of Teacher Education, Certification and Inservice Staff Development, serves a coordinating and articulating function, by developing and implementing registration procedures noting colleges which graduated the beginners, programs pursued, and beginning and finishing dates in the Beginning Teacher Program.

The BTP can and should serve as an axis of professional preparation, providing performance data on the products of colleges as they evaluate their preservice programs, and to districts as they strive to improve their inservice programs. It goes without saying, however, that the BTP must itself be of high quality if it is to play this central role.

The BTP is now in its third year of operation and has already produced noticeable effects. Its positive influence on college programs and inservice district programs can be documented. However, influence must flow in two ways. The advantages that the BTP can gain from utilization of college instructors and district personnel in support activities have barely been tapped. This lack of adequate cooperation among these components is only one example of weaknesses that require attention if the program's potential benefits are to be realized.

The major problems of the BTP cannot be overcome by acting on a few quick decisions. They are far too complex for that. Whatever moves are

made to improve the program should be based upon research and thorough policy deliberations. With this warning, the following recommendations are made:

1. Research Studies

   a) Studies should be made to set a minimum standard of performance of the essential generic competencies.

      Justification: If a state is to grant a person the privilege of being employed as a teacher, it should make as certain as possible that the individual is competent. This principle dictates that a minimum standard of competence be clearly specified in terms of the Florida essential competencies.

   b) A study of the portfolio system should be made, including content and the relationship between the portfolio and the beginning teacher's success as measured by summative observations.

      Justification: The development and maintenance of portfolios is a time-consuming task. For this reason alone, its usefulness and the benefits occuring from it should be verified by objective study.

   c) A study of the effectiveness of inservice activities and materials used in BTP should be conducted.

Justification: The assistance given to the beginning teacher represents inservice dollars and staff time. This fact alone indicates the need to assess training activities and materials. Information gained from such a study should be useful to both district inservice programs and Colleges of Education.

d) Study of the impact of BTP upon the performance of the beginning teacher should be continued.

Justification: That the behavior of the beginning teacher changes during the first year is apparent from a preliminary study reported herein, but it is not known whether the change is attributable to the program of support or simply to the year's experience normally undergone by teachers in general. Determination of the program's worth is contingent upon an answer to this question.

e) A study of the use of funds now allocated to the BTP should be undertaken.

Justification: The BTP is funded at the rate of $1.70 per unweighted full-time equivalent student in the school district. A study of the use of these funds should provide information to enable the Department of Education to recommend any changes needed in the funding procedure.

f) The state legislature has enacted legislation establishing an alternative teacher preparation program that enables a four

year graduate of an accredited college to teach while undergoing training by a school district. In view of this alternative, it is appropriate to ask whether or not such a program of preparation produces competent teachers. To answer this important question, research should be undertaken to ascertain (1) the type and extent of preparation provided by the districts and how it compares with the BTP; (2) how the performance of the alternative teachers compares with that of the beginning teachers as measured by the summative instruments of the FPMS; and (3) the effect of the alternative program as determined by the criteria specified in the legislation.

2. Policy Studies

a) Assuming that observations of classroom performance should be made by trained observers, who should these observers be and by whom should they be trained?

b) By what criteria should observers be certified and for what period should the certificate be valid?

c) How many summative observations should be required for each beginner and by how many different observers?

d) Should legislation or State Board Rule be sought that would enable the Department of Education to require submission of all observations made with summative instruments?

e) What should be the role of associate master and master teachers in servicing the BTP?

f) How many formative observations should be made of each beginning teacher? Should the teacher's competence be judged by performance in each domain or by summative observations alone?

g) What preparation should be provided for the support staff to enable it to function effectively?

h) What should be the role of college personnel in the BTP?

Chapter II

Knowledge Base for Teaching Higher Order Thinking

The ability to solve problems is one of the perennial objectives of schooling, emphasized by the profession and the public alike. That students may master the content of the school subjects and yet remain unable to cope with problems of daily living or in the world of work is well known. This is because subject matter can be mastered without learning the mental skills -- variously referred to as intellectual skills, problem-solving skills, thinking skills, higher mental processes, critical thinking, and reasoning -- that problem solving requires.

Moreover, subject matter preparation provides no assurance that teachers themselves thereby acquire problem-solving skills. This fact is not understood by lay persons. According to the popular view, anyone who has acquired the information comprising liberal arts education, and certainly a master's degree in liberal arts, has sufficient knowledge to teach intellectual skills as well as subject matter. But this is not the case. To be able to recall academic knowledge, to comprehend it, to do the exercises required in college courses is all good and well. Nevertheless, teachers must know more about subject matter than that; they must be able to analyze and present the content of instruction in challenging and thought provoking ways. That entails ability to understand and handle the logical relations of ideas in a context of problem definition and solution. For example, it is one thing to require students to learn definitions; it is quite another thing to know

what a definition is, to know its parts, its truth status, and to be able to explain these essentials to students. To understand what a statement says is not to know whether to accept it, whether it is logically fallacious, whether s/he who utters the statement stands to profit from it, or to know what facts are relevant. Indeed, there is considerable evidence that many teachers do not possess such knowledge sufficiently well to teach it, although they may have more academic work than the equivalent of a major in liberal arts.

The foregoing considerations alone are enough to justify greater emphasis on intellectual skills in teachers' professional preparation and in the school's program. But there is still further reason. The density of social events and the rapidity with which occupations come and go require flexibility and modes of adjustment best made through the exercise of higher mental processes.

While the teaching profession has long claimed intellectual skills as one of its objectives, past efforts to provide for this objective have been sporadic and feeble. Nevertheless, past efforts to include instruction in these basic skills are being revived. Connecticut and California are including these skills in their state testing programs, and a number of other states are considering this policy. Furthermore, a number of school districts throughout the nation -- some in Florida -- are offering courses in these intellectual skills.

In the fall of 1984, the Teacher Evaluation Project, under the auspices of the Florida Coalition for the Development of a Performance Measurement System, began a literature search for research knowledge on

the higher mental processes. We are assembling this research and developing plans for incorporating intellectual skills in the school program and for the development and measurement of teacher performance in these skills.

Fortunately there is an extensive body of research on teaching and learning higher mental processes. This body of research has now developed to the point that it can be digested and organized into a pedagogical program involving teacher performance in the handling of instructional content. Those of us who are working on this project are now organizing the results of this research into domains of knowledge for the development and evaluation of teacher performance as it relates to teaching problem-solving skills.

We have assembled a bibliography of some 400 research studies. From our analysis of these studies, it is clear that there are four kinds of problems, each having its own characteristics and requiring its own set of intellectual skills for solution. These problems are:

1. Partly-structured problems, identified as perplexities where known concepts, principles, and procedures are not sufficient to generate a solution and where criteria for deciding among possible solutions are debatable.
2. Classification problems, identified as those that require the formulation and testing of criteria for grouping objects, events, and so forth.
3. Evaluation problems, identified as situations where a person is required to assess what someone expresses as a conclusion, or sets forth as an argument.

4. Well-structured problems, identified as problems for which there are agreed-upon rules, formulas, and procedures for solving them and criteria for telling when they are solved.

Each type of problem entails teaching skills for handling the content of instruction and directing learning activities. These skills can be grouped into categories of teacher behavior for which indicators can be identified and described. We refer to each type of problem as a domain of teacher behavior, and the categories of instructional skills for each problem type as concepts, and descriptors of skills as indicators. A few examples of each of these elements will add a bit of concreteness.

Domain 1.0: Partly-Structured Problems

Definition: This domain = perplexities where known concepts, principles, and procedures are not sufficient to generate a solution and where criteria for deciding among probable solutions are debatable.

These activities are grouped under the following concepts:

1.1 Concept: Problem State

Definition: Problem state = events or states of affairs incongruous with what would typically be expected; a baffling or seemingly inexplicable situation; a discrepant event.

1.1.1 Sample Indicator: Presents Means-ends Discrepancy

Definition: This indicator = teacher demonstrates or describes a perplexity for which available means are inadequate to obtain indicated ends,

or teacher induces students to do so.

1.2 Concept: Situation Analysis

Definition: Situation analysis = search for details in order to identify variables that appear to be related to the resolution of a discrepancy or an enigma.

1.2.1 Sample Indicator: Emphasizes First Observation as Data Gathering

Definition: This indicator = teacher instructs students to use their first observation to analyze the situation rather than to judge hunches about solutions.

1.3 Concept: Flexibility

Definition: Flexibility = ability to keep options open, to evolve novel approaches to the resolution of perplexities.

1.3.1 Sample Indicator: Uses Groups

Definition: This indicator = teacher creates groups wherein student interaction can trigger new hypotheses.

1.4 Concept: Variable Search

Definition: Variable search = engagement in activities whose end is to identify, analyze, and verbalize variables.

1.4.1 Sample Indicator: Confounds Variables

Definition: This indicator = teacher suggests variables that challenge students with alternatives to their proposed solutions or hypotheses.

1.5 Concept: Hypothesis Evaluation

Definition: Hypothesis evaluation = examination of proposed solutions to decide their testability and which one(s) is most likely to be valid.

1.5.1 Sample Indicator: Clarifies Hypotheses

Definition: This indicator = teacher involves students in the task of formulating proposed solutions in a clear "if-then" form.

1.6 Concept: Variable Control

Definition: Variable control = development of designs and techniques for testing proposed solutions.

1.6.1 Sample Indicator: Search for Ways to Falsify Hypotheses

Definition: This indicator = teacher challenges students to seek tests or varied conditions that disconfirm, as well as tests that confirm, proposed solutions.

1.7 Concept: Correlation of Variables

Definition: Correlation of variables = activities that involve recognition and interpretation of relationships among variables that randomly or otherwise fluctuate.

1.7.1 Sample Indicator: Correlation Situation

Definition: This indicator = provision of a situation in which there are two sets of variables and their relationship is to be estimated by scattergrams.

Domain 2.0: Classification Problem

Definition: This domain = perplexities whose solutions require the ordering of things -- objects, symbols, events, ideas, sets -- into groups or classes.

These activities are grouped under the following concepts:

2.1 Concept: Conjunctive Classification

Definition: Conjunctive classification = categorizing things -- objects, events, etc. -- in groups or classes through the use of one or more attribute jointly.

2.1.1 Sample Indicator: Develop Conjunctive Classes

Definition: This indicator = teacher behavior that encourages students to develop and relate their own criteria, schema, or rules for classifying.

2.2 Concept: Disjunctive Classification

Definition: Disjunctive classification = categorizing things -- objects, events, etc. -- into groups or classes based on alternative attributes or criteria.

2.2.1 Sample Indicator: Develop Disjunctive Classes

Definition: This indicator = teacher involves students in categorizing things -- objects, events, etc. -- into groups where they are required to develop alternative criteria or to identify alternative classifying attributes.

Domain 3.0: Evaluation Problems

Definition: This domain = perplexities whose resolution requires processes of evaluating what someone expresses as a claim or

conclusion, or sets forth as an argument.

These activities are grouped under the following concepts:

3.1 Concept: Parts of an Argument

Definition: Parts of an argument = statements -- definitions, principles, etc. -- that comprise an argument.

3.1.1 Identifies Conclusion

Definition: This indicator = teacher behavior that directs students in the task of recognizing and stating the conclusion of a chain of reasoning.

3.1.2 Identifies Premises

Definition: This indicator = teacher behavior that directs students in the task of recognizing and expressing the major and minor statements on which a conclusion rests.

3.1.3 Identifies Key Words

Definition: This indicator = teacher performance that engages students in the task of picking out words or expressions on whose meaning an argument hinges.

3.1.4 Recognizes Word Vagueness or Ambiguity

Definition: This indicator = teacher performance that engages students in explicating emotionally charged words, slogans, symbols, and equivocations.

3.1.5 Identifies Parts of Definition

Definition: This indicator = teacher behavior that encourages students to analyze definitions into (a) the term or expression to be defined, and (b)

words or expressions that define it, and to identify the referent when possible.

3.2 Concept: Deductive Argument

Definition: Deductive argument = course of reasoning whose validity rests on logic alone, and that advances from premises to conclusion.

3.2.1 Evaluates Deductive Reasoning

Definition: This indicator = teacher performance that provides experience in examining arguments where validity of the conclusion depends entirely on logical relations between premises (assumptions) and conclusions.

3.2.2 Distinguishes Necessary and Sufficient Conditions

Definition: This indicator = teacher behavior that provides experience in recognizing the condition necessary for something to happen, and whether or not is is sufficient for it to occur.

3.3 Concept: Inductive Argument

Definition: Inductive argument = course of reasoning based upon probable inference and the truth of whose conclusion is probable.

3.3.1 Identifies Probable Conclusions or Inferences

Definition: This indicator = teacher performance that provides experience in recognizing conclusions based upon partial evidence.

3.3.2 Evaluates Explanations

Definition: This indicator = teacher behavior that

engages students in the task of deciding whether or not a particular explanation is justified by the data, by its consistency with other evidence, and is simpler than other proposed explanations.

3.3.3 Identifies Overgeneralized Explanations or Conclusions

Definition: This indicator = teacher behavior that engages students in recognizing flaws in reasoning attributable to sampling error.

3.4 Concept: Evaluation of Information

Definition: Evaluation of information = judgment of the reliability and validity of information.

3.4.1 Evaluates Authoritative Information

Definition: This indicator = teacher behavior that provides experience in judging information by the following criteria: a) the person providing the information has no conflict of interest; b) the person's experience and expertness covers the area from which the information is given; c) the information is based on observation rather than inference; d) the information is not contrary to what we have good reason to believe.

3.4.2 Evaluates Observation Information

Definition: This indicator = teacher performance that provides experience in judging the validity of information by the following criteria: a) the information is taken from a written record made near the time of the observation, and b) the information

was gained from high quality techniques, skillfully used.

3.5 Concept: Common Flaws

Definition: Common flaws = common fallacies in reasoning.

3.5.1 Identifies Black-or-White Fallacy

Definition: This indicator = teacher performance that provides experience in picking out arguments where one alternative is emphasized to the neglect of the other(s) or to the disregard of intermediate positions or different courses of action.

3.5.2 Identifies Fallacy of Oversimplification

Definition: This indicator = teacher activities that engage students in the practice of recognizing explanations that are too simple for the events or causes.

3.5.3 Identifies Post Hoc Fallacy

Definition: This indicator = teacher behavior that encourages students to recognize explanations whose claim to validity rests on mere temporal succession.

Domain 4.0: Well-Structured Problems

Definition: This domain = perplexities for which there are agreed-upon rules, formulas, or procedures for solving them, and criteria for telling when they are solved.

These activities are grouped under the following concepts:

4.1 Concept: Problem Structure

Definition: Problem structure = a formulated problem with one or more rules or procedures for solving it, and criteria for

telling whether a trial solution is correct.

4.1.1 Sample Indicator: States Problem in Clear Familiar Language

Definition: This indicator = teacher behavior that presents problem statement in terms familiar to or defined for the students.

4.2 Concept: Explanatory Principles, Rules, and Procedures

Definition: Explanatory principles, rules, procedures = guidelines for solving problems of a specified type.

4.2.1 Sample Indicator: Encourages Graphic Representation of Problem Parts

Definition: This indicator = teacher behavior that encourages students to outline or depict problem elements in ways that make their relationships clear and lead to correct selection and application of rules and procedures: using graphs, diagrams, etc.

4.3 Concept: Application

Definition: Application = application of principles in situations that range from application of rule or procedure to problems involving identical materials to analogous problems, and finally to situations not specifically similar.

4.3.1 Sample indicator: Presents Application Situation

Definition: This indicator = teacher behavior that supplies analogous or novel situations to which the learned principle or procedure is to be applied.

4.3.2 Sample indicator: Specifies How Similar Tasks Differ

Definition: This indicator = teacher behavior that points

out those elements of the new task which will call for differences in the nature or order of procedures in the solution of the problem.

4.4 Concept: Verification Procedures

Definition: Verification procedures = a procedure by which to judge whether or not a solution is correct.

4.4.1 Sample Indicator: Compares Solution with Problem Demands

Definition: This indicator = teacher behavior that induces the student to determine whether the solution reached is the one called for in the problem.

Recommendations

To complete the task of assembling the research and refining the concepts and indicators, and to develop training materials and instruments for classroom observation to measure teacher performance in the teaching of intellectual skills and procedures, it is recommended that provisions be made to:

1. Complete compilation of the research findings.
2. Continue work on the extension and refinement of the concepts and indicators of effective teaching practices.
3. Develop formative and summative observation instruments.
4. Develop a manual for teachers, setting forth the essential knowledge about intellectual skills.

Chapter III

Development of the Conferential Observation Instrument

Principals, supervisors, peer teachers or any other educators working with the beginning teacher must communicate the results of an observation in an effective manner if the information is to have a positive impact on the classroom performance of the teacher. Skills of effective communication that foster respect, trust, and confidence enhance the possibility that the teacher will accept and fully participate in the planning and activities for his or her improvement. It is the observer's responsibility to conduct the conference and to skillfully guide the teacher through consideration of the observation data, formulation of the problem(s), and development of plans to solve the problem(s). The observer must understand the conferring process and he/she must have the pedagogical knowledge and skills to provide specific assistance to the teacher.

The trainers in the Beginning Teacher Program have repeatedly asked for help in training personnel in conferential skills. According to these trainers, a weak point in the procedures used to implement the Florida Performance Measurement System is the communication of the results of the beginning teacher observations to the teacher by the school administrators, the peer teacher, or other professional personnel. Interviews with a sample of some 150 beginners bear out this opinion.

Principals and others have been trained and tested in the skills of observation. They now need to be trained in the skills of conferring.

This persistent need has led us to search the literature on conferring and to develop an observation instrument based on available research; an instrument for training principals and other personnel in the skills of conferring where the basis of the conference is data derived from observation of classroom performance.

For convenience we used the term "consultant" to designate the principal, supervisor, classroom teacher (designated as peer teacher in the Beginning Teacher Program), or anyone who provides help to a teacher whose need is decided on the basis of data derived from observations of classroom performance.

Development of Knowledge Base

The knowledge and skills needed by a consultant were identified by reviewing the research on consultation. From a survey of the literature, we identified four major concepts: (a) the development of rapport, (b) interpersonal skills, (c) problem formulation, and (d) development of plans for solving the problem and for the teacher's improvement. These concepts were expanded into a total of 25 indicators, each of which was defined and exemplified. Each concept was followed by a summary of the research evidence bearing upon its use and efficacy in the conferring situation.

Development of Instrument and Coding Manual

The concepts and indicators derived from the research literature were incorporated into an observation instrument. The instrument provides for the coding of both effective and ineffective consultant behavior. It requires that the observer tally the overt behavior of the consultant without judging the quality of the performance. The data are analyzed and interpreted apart from the observation to indicate conferential skills that the consultant needs to improve.

Several videotapes have been prepared for training with this instrument, and criterion tapes are used to test for observer accuracy. (Appendix B contains a copy of the instrument, and Teacher Evaluation and Assessment Center Report lists the training and criterion tapes.)

A manual for coding consultant performance was prepared as a guide for those who prepare trainers and observers. It points out salient features of each indicator and gives examples. Terms such as "uses specific reinforcement" are exemplified. Specific reinforcement occurs when the consultant specifies the behavior being praised, for example, "You do a good job of following through on your classroom rules by commenting on the examples of compliance as well as examples of noncompliance". The observer learns to distinguish specific reinforcement from general reinforcement. In relation to general praise, the manual says "gives praise without referring to specific behavior", for example, "You are doing beautifully," is an instance of general reinforcement.

The coding manual's numbered indicators correspond to the numbered indicators on the instrument and provide useful training material to bring observers closer together in matching behavior with the behavioral indicators during practice.

Pilot Studies of the Materials

Thirteen teachers and administrators from Florida's Panhandle participated in a split-half field test of the observation instrument. The primary purpose of the field test was to determine instrument codability. This entailed, among other things, an examination of the completeness and accuracy of research documentation, clarity and specificity of the language throughout the materials, and codability of the videotapes using the coding manual and instrument.

As a result of the first field test, the materials were revised and edited for use in the second field test. The language was changed to simplify items on the instrument and two items were combined -- "elicits teacher's ideas" was added to "explores planned alternatives", thereby reducing the total number of effective indicators from 20 to 19. On the ineffective side of the instrument "does not follow up" was dropped because it is evident from its omission on the positive side, thereby reducing the number of ineffective indicators from 20 to 19 also.

During January of 1985, 15 teachers and administrators participated in a second field test of the revised Domain 10.0 materials. Participants read and critiqued the domain document and coding manual for conferring for clarity of content, and then coded two videotapes

using the conferring instrument. After coding the videotapes, the participants discussed their coding and noted items where coding differences existed. The participants then coded two criterion videotapes.

As a result of the second field test, the instrument was simplified by reducing the number of items to fifteen on each side. Not all ineffective items are opposites of effective items, so there are eighteen numbered spaces on each side of the instrument. Since so many changes in the instrument occurred as a result of the first field test, no reliability estimates were computed.

The manual for coding this instrument was revised to reflect the changes on the revised instrument. Some of the examples were rewritten, notes were added, and some terminology was clarified. The revised manual, dated June 1985, and the Domain 10.0 document are available from Dr. Garfield Wilson, Office of Teacher Education, Certification, Staff Development and Professional Practice Services, Department of Education, Tallahassee, Florida.

Recommendations

1. Development and testing of training programs in the conferential skills that would be available to master trainers so that they can help districts prepare their personnel in conferring.
2. Development of additional training videotapes for practice during training.

3. Development of procedures for certifying peer teachers and administrators in conferential skills. This requires: a) development of cognitive examinations on the knowledge base, and b) development of criteria videotapes for testing observer accuracy.

Chapter IV

Predictive Validity Studies: Studies Conducted to Test the Relationship Between Teacher Performance as Measured by the FPMS and Measures of Student Outcome

Background

Developmental work on the Florida Performance Measurement System began in 1981. By the end of 1982-1983, the teacher effectiveness research literature had been assembled into five domains and reviewed by nationally recognized experts. A sixth domain was developed in 1983-1984. These domains with their concepts and indicators represent the knowledge base of the Florida Performance Measurement System (FPMS).

Observation instruments were developed and field tested simultaneously with the development of the knowledge base. Preliminary testing for validity and extensive testing for reliability were conducted on the FPMS summative instrument during 1982-1983 (FPMS Reliability Study, 1983). A quartile-based scoring approach produced reliability estimates of r=.91 for the effective behaviors and r=.84 for the ineffective behaviors.

The FPMS summative instrument consists of 20 items designated effective teaching and 19 ineffective behavioral indicators. Each indicator on the summative instrument is a construct of one or more observable behaviors that can be traced to a formative instrument and a domain of research.

During 1983-84, 117 trained and certified observers collected data from over 40 schools in 11 school districts using a stratified cluster sampling technique. Strata included geographical location, population density, grade levels, and socio-economic status. Demographic data were obtained on numerous context/frame factor variables of teachers and students.

Results from the FPMS Norming Study (1984) suggested that the only difference between groupings of primary interest was a substantial difference in scores between elementary (K-5) and post-elementary (6-12) teachers. Furthermore, the data established that the FPMS summative instrument met the following measurement criteria -- content validity, reliability, feasibility, scalibility, flexibility, discrimination, feedback, and relevance.

The products of the research and development work briefly described above were carried forward as the basis for the initiation of predictive validity studies on the FPMS summative instrument.

1984-1985 Predictive Validity Studies

Norming studies are conducted on representative samples of teachers and schools to determine the influence of variation in teaching contexts on scores. One norming study was conducted for the FPMS. By contrast, predictive validity studies are situation specific. Multiple studies that assess teacher behavior in a variety of grade levels, subject areas, lengths of time and various outcome measures are being conducted to estimate predictive/criterion validity of the FPMS.

Data for both norming and predictive validity studies must be collected via observation during regular class periods. Classrooms are the researcher's laboratory for these studies. Three school districts cooperated in the 1984-1985 studies. In addition, a Department of Education sponsored project conducted by the Office of Clinical Education at the University of South Florida participated. Table 12 gives a summary description of these studies.

TABLE 12

Summary Outline of 1984-1985 Predictive Validity Studies*

Name	Subject	Grades	Number of Classrooms	Length	Pretest	Posttest
Study A	Math	1, 3, 5	21	2 wks.	CTBS/ Special Subject	Special Subject
	Social Studies	1, 3, 5	31	2 wks.	CTBS/ Special Subject	Special Subject
Study B	American History	8, 11	40	16 wks.	CAT	Special Subject
Study C	Multiple Subject	6	35	32 wks.	CAP	CAP
Study D	American Literature	11	35	32 wks.	CAP	CAP
Study E	Typing	9-12	40	9 wks.	CAT/ Skill	Skill

*At the time of this report, data collection and analysis for Studies A and B had been completed and data collection was in progress for Studies C, D, and E.

Research Questions

The studies conducted for purposes of testing the predictive validity of the FPMS summative instrument are primarily concerned with this research question:

1. Do FPMS scores relate positively with measures of student outcome? Two different types of outcome measures were investigated to answer this question:
   a) Task engagement - a measure of student learning activities defined as the average percentage of students engaged in teacher specified activities, and
   b) A variety of achievement tests.

In addition, two subordinate questions were investigated, as follows:

1. Is the FPMS quartile-based score an appropriate scaling for this instrument?
2. Are there other scalings, or methods of treating various items on the instrument, that relate either more positively or negatively with student outcome measures than a quartile-based total score?

At the time this report was prepared, only studies A and B were completed. The remainder of this report consists of these two studies. To answer these questions, several studies must be conducted.

Study A

Predictive Validity Studies in Elementary Mathematics and Social Studies

Two predictive validity studies on teacher interns were conducted by the Office of Clinical Education (OCE) at the University of South Florida as part of a Department of Education (DOE) grant. The original purpose was to determine the effects of training interns in FPMS teacher behaviors on classroom performance and student achievement (Department of Education Study, 1985). The data were subsequently reanalyzed for purposes of the present study.

Sample

Two subject matter areas -- math and social studies -- were represented with 21 classes in math and 31 in social studies, or 52 classes with 1100 students. Classes were conducted in grades two, three, and five in Hillsborough and Pinellas districts.

The two subject areas were treated both as a whole group and separately in the analyses. Treatment as a single sample only (n=52) may increase the power of tests of significance, but it could confound the relationship between FPMS scores and measures of student outcome. The relationship between FPMS scores and achievement could differ from subject to subject, as could the relationship between FPMS scores and task engagement.

The large number comprising this sample makes it possible to compare the total sample with those obtained for specific teachers. Despite the lack of random assignment and selection, one may consider the total sample to be representative of the populations found in the geographic areas under study due to the large sample size and varied sources of subjects. The same, of course, cannot be said of the samples found in specific classes. Upon close inspection, some interesting phenomena became apparent:

1. Some classes lacked any students scoring in the top 50 percent of all students on the pretest.
2. Some classes lacked any students scoring in the bottom 50 percent of all students on the pretest.
3. And, amazingly enough, some classes lacked any students scoring in the middle 50 percent of all students on the pretest.

It is obvious from the above that serious biases exist in the samples drawn for specific classes, clearly showing that students in these studies are not randomly assigned to classes.

Design and Procedure

The summative instrument of the FPMS consists of 39 items, 20 indicators of effective teaching and 19 indicators of ineffective teaching. All items are scaled positively; that is, more behaviors result in higher scores for the effective items, and fewer behaviors result in higher scores for the ineffective items.

The study was conducted over a two-week period. Teacher interns were given a unit to teach during this period, and tests were given before the beginning of the unit and at the end of the unit. Interns did not receive copies of the test.

The Teacher Education Internship Project final report (1984) describes the locally developed tests:

> Six unit tests, one for each of the instructional units, were developed and pilot tested. Each test had two parallel forms. One form was used as a pre-test, the other as the post-test. Tests were developed based on the content of the units and the learner objectives identified.
>
> The Grade 2 Social Studies test contained 18 items, the Grade 3 and Grade 5 tests contained 22 items. Each of these tests contained true/false and multiple choice items. The Mathematics tests contained multiple choice and completion-type questions in which the student was asked to supply the correct answer. The tests at all three grade levels were 20-item tests. The maximum score for the Grade 2 and Grade 3 tests was 20, while the maximum score for the Grade 5 test was 27. The tests were reviewed to assess their content validity and pilot tested. Data obtained from the pilot testing were used to make revisions where necessary.

The means and standard deviations across grade levels were almost identical on both pre and post-tests. Neither a ceiling nor a floor effect was apparent on these tests, and distributions exhibited neither unusual skewness or kurtosis (unusual being defined as .30 or above).

The interns were observed three times. One observation was conducted prior to the two-week lessons, and was therefore not included in the analyses. Two observations were conducted during the two-week lesson period, and these were averaged to obtain performance scores for each intern on the FPMS summative instrument.

Each intern was observed by two trained and certified FPMS observers. In addition, other trained observers conducted spot checks on classes, and their observations were compared to the regular observers. Only small differences were found, providing support for the consistency and validity of observations.

Selection of appropriate dependent variables is fundamental to the success of predictive validity studies. For example, people commonly hold teachers directly responsible for student achievement scores. This practice is questionable, for individual teachers have an important but very limited influence (estimated by some to be between 5 percent and 20 percent) on what students learn in a given class (McDonald, 1976).

Medley (1985) outlined a model for examining teacher performance in relation to teacher effectiveness -- the essence of predictive validity studies. His model depicts a progression from teacher performance to student learning experiences and from student learning experiences to measures of student outcome such as scores on achievement tests. One measure of learning experience is the percent of students on task as coded by observers during regular FPMS observations. This measure, hereafter referred to as task engagement, was introduced in this study.

From a student's perspective, engagement on the task assigned by the teacher may be important for a higher test score. Thus, a relationship between student engagement and student test scores may be expected. From the teacher's perspective, causing students to listen during discussions and to do assigned seatwork may have a direct and positive effect on their test scores. Thus task engagement was examined as a

potential mediating variable between teacher behaviors and student test scores. Although achievement scores are also of particular interest in studies of this nature, one expects stronger relationships to exist between teacher behavior and student activities such as task engagement than between teacher behaviors and achievement test scores.

Task engagement for purposes of this study, was defined as the number of students on task as determined by observers at selected times during the regular class period. Four estimates of task engagement were obtained for each observation and averaged to produce an estimate of general engagement for class. Task engagement was coded concurrently with the coding of teacher behavior on the FPMS summative instrument. Information on frame factor and instruction format variables was also collected, including grade level, subject area, teacher's perception of students' SES level, free lunch data for schools, grouping and method of instruction.

School-based free lunch data obtained from DOE reports provided the main estimate of Socio-Economic Status (SES). This variable was inverted, so that higher scores meant higher SES and lower scores, lower SES.

Analyses

Both Spearman (non-parametric) and Pearson (parametric) correlations were run. Since they produced nearly identical results, only Pearson correlations are cited in these studies. Two scalings of observation data from the FPMS summative instrument, the quartile-based scaling

developed during the 1983-1984 Norming Study and the Blom expected normal score scaling (ENS) were applied with similar results. Unless otherwise specified, all analyses discussed in the remainder of this report were run on the class mean raw posttest scores for the average stratum. This provides an estimate of teacher effects with average students.

Results

FPMS quartile-based scale results are shown in this section for the total FPMS scale, comprising 39 items corresponding to both effective and ineffective behaviors. This is the scaling used for Master Teacher ranking.

To address the question of scaling effectiveness, proposed ealier, two possible scalings are included in Table 13, the quartile-based score and the Blom expected normal score (ENS). This table shows that any advantage gained by using the ENS scaling for the math task engagement variable is more than offset by the advantage shown by the quartile-based scaling for the social studies classes. This general similarity was found for all correlations, thus, emphasis is placed on the quartile-based scaling throughout the remainder of this report.

It is also obvious from Table 13 that the FPMS quartile-based score exhibits a strong correlation with student task engagement for both math (r=.57) and social studies classes (r=.47), and a moderate correlation with student achievement for both math (r=.31) and social studies (r=.28). These results provide strong support for the use of the quartile-based FPMS total score as a measure of teacher effectiveness.

TABLE 13

Correlations Between FPMS Scores and Outcomes

Average Academic Stratum

		Math			Social Studies		
		Task Engage n=17	Gain Score n=17	Post Test n=17	Task Engage n=30	Gain Score n=30	Post Test n=30
Quartile Total Score	r=	.57	.31	.31	.47	.27	.28
	P=	.009	.11	.11	.005	.08	.06
ENS Total Score	r=	.69	.33	.31	.36	.17	.20
	P=	.001	.10	.11	.03	.19	.15

TABLE 14

Correlations Between Task Engagement and FPMS Scores - Math

		Task Engagement			
		All Students n=21	High Achievers n=15	Average Achievers n=17	Low Achievers n=20
Quartile Total Score	r=	.42	.60	.57	.49
Effective Subscore	r=	.17	.35	.32	.26
Ineffective Subscore	r=	.76	.80	.78	.76

for n = 15
P = .10 at r = .44
P = .05 at r = .51
P = .01 at r = .64

for n = 20
P = .10 at r = .38
P = .05 at r = .44
P = .01 at r = .56

The ineffective subscore exhibits a highly significant relationship with task engagment (.01), while the effective subscore correlations

border on significance. There is a significant difference between the correlations of effective and ineffective subscores and task engagement (Tables 14 and 15).

TABLE 15

Correlations Between Task Engagement and FPMS Scores-Soc Std

			Task Engagement		
		All Students n=31	High Achievers n=29	Average Achievers n=30	Low Achievers n=30
Quartile Total Score	r=	.44	.51	.47	.47
Effective Subscore	r=	.28	.36	.31	.31
Ineffective Subscore	r=	.58	.59	.58	.58

for n = 30
P = .10 at r = .31
P = .05 at r = .36
P = .01 at r = .46

Correlations for the average achievement stratum of students are r=.28 for social studies and r=.31 for math (Tables 16 and 17). For higher and lower achieving students, the relationships are greater in math (r=.45) than for social studies (r=.10).

There were no significant differences between FPMS scores and frame factors or instruction format variables. For example, the relationship between FPMS scores and socio-economic status estimates bordered on zero (r= .09), and there were no significant differences between the three grade levels for FPMS scores.

TABLE 16

Correlations Between FPMS Scores and Posttests - Math

	All Students n=21	High Achievers n=15	Average Achievers n=17	Low Achievers n=20
Quartile Effective Score	.23	.31	.18	.26
Quartile Ineffective Score	.41	.54	.42	.44
Quartile Total Score	.33	.45	.31	.37

for n = 15
P = .10 at r = .44
P = .05 at r = .51
P = .01 at r = .64

for n = 20
P = .10 at r = .38
P = .05 at r = .44
P = .01 at r = .56

TABLE 17

Correlations Between FPMS Scores and Posttests - Social Std

	All Students n=31	High Achievers n=29	Average Achievers n=30	Low Achievers n=30
Quartile Effective Score	.16	.14	.32	.20
Quartile Ineffective Score	.02	.08	.06	.02
Quartile Total Score	.13	.09	.28	.17

for n = 30
P = .10 at r = .31
P = .05 at r = .36
P = .01 at r = .46

Criterion Comparisons

An additional test was conducted to determine how the teacher interns observed in this study would be distributed according to the norms developed in the 1984 Norming Study.

Table 18 shows that 32 of the 46 teachers scored below the mean (x) of the Norming Study. This was expected since these were inexperienced teacher interns. It is important to note, however, that the average students in top teachers' classes scored higher (although not significantly higher) than average students in bottom teachers' classes.

TABLE 18

Comparison of Higher and Lower Scoring Teachers -- Study A

Achievement

Elementary Percentile Ranks	Class Mean Raw Posttest Scores for Middle 50% of all Students			
	N	Mean	z	Prob z
Teachers Above 50th Percentile	14	17.26	1.53	.13
Teachers Below 50th Percentile	32	16.05		

The results of such comparisons are somewhat muddied by the teachers in the middle of the distribution, who, although scoring very nearly the same on the FPMS instrument, are arbitrarily assigned either a top or bottom designation. For example, significance tests, comparing average

task engagement for the top against bottom groups, proved non-significant despite the sizable and significant correlation between this variable and FPMS scores (Table 19).

TABLE 19

Comparison of Higher and Lower Scoring Teachers -- Study A

Task Engagement

	Higher Teachers	Lower Teachers
Number of Cases	14	33
Mean	.89	.86
Probabilities of significant difference	.29	

The nonparametric Mann-Whitney U (Wilcoxon Rank-Sum) test of significance was run upon the data. Due to the small sample size, and the precise probability specification resulting from this test, it is best for meta-analytic uses reported later in this paper.

Discussion and Implications

Results of this study indicate that the relationship between task engagement and achievement is positive and highly significant for both subject areas, and that the FPMS summative instrument correlates positively with student achievement measures for both math and social studies classes using either an expected normal score scaling (ENS), or its present quartile-based scaling.

The results in Study A came from classes taught by teacher interns. According to our data, experienced teachers rarely exhibit as many ineffective behaviors as do beginning teachers. This may have an effect on the validity coefficients of the FPMS instrument in future studies, particularly on the ineffective side of the instrument.

The content taught and examined in these studies was extremely specific, and the time periods very short. It is important to remember that these results are suggestive, but not generalizeable. Significance levels should be viewed with caution due to the multiplicity of the several analyses. These are considered exploratory studies. However, the general consistency of correlations makes a strong case for the use of the probabilities associated with median correlations as an estimate of significance.

It is difficult to separate the effects of the regular classroom teacher from those of the interns for this study, as the interns only dealt with students part time. However, only the interns taught the lessons used in this study and the achievement test was designed to specifically measure the knowledge gained in lessons taught by the interns. The strong relationship between engagement and FPMS scores, however, does not suffer much from this limitation, as it was obtained concurrently with the observation score.

The highly specific content area, the formalized nature of task materials, the short time involved in the study and the inexperienced nature of the teachers all could reduce the effects of ambient content factors and increase the effects of task engagement within this sample.

In general, a student's involvement in activities directly related to test items should improve his/her score. This factor may not prove as important over longer periods and for more complex tasks.

Study B

Predictive Validity Study in Secondary American History

The main purpose of this study was to determine the relationship between FPMS summative instrument scores and student outcomes in post-elementary American history classes.

Sample

This study was conducted in a southern Florida school district. An attempt was made to obtain data from all practicing middle and secondary school classes teaching a specific American history course. Results were obtained for a total of 40 teachers in the 8th and 11th grades. Teacher experience ranged from two years to 37 years. Participation was voluntary, however, only one teacher declined participation.

Test score results were obtained from almost all middle and high school students enrolled in the American history classes during the fall semester 1984 (n=1050).

Design and Procedure

This is a process-product study where the process is defined as teacher behavior as coded on the FPMS summative instrument (independent variables) and the products are defined as measures of student task engagement and scores on a subject specific, standardized achievement test (dependent variables).

Coding of the summative instrument was conducted by observers who were trained and certified in the FPMS summative observation instrument using state level certification standards. Coding of student task engagement was conducted concurrently with coding of the summative instrument. Observers received separate training for this purpose.

Two observations of each teacher by two observers on two occasions were obtained for each teacher during the fall semester 1984. Four different estimates of task engagement were taken during each observation and then averaged to obtained this measure. Four observers conducted all of the observations, and these were selected as top scorers on the FPMS certification tests.

A county-wide, standardized American history test was used as a posttest. The CAT General Reading percentile ranks were used as a covariate. The history tests were county-developed tests normed for the grade levels. Raw scores ranged from 0 to 35. The tests were developed to test the objectives defined for American history courses at the specific grade levels. The history test was administered by each teacher on or about the last day of the semester.

Independent variables of interest included all item and subscores of the FPMS. Scoring of ineffective FPMS items was reversed, as it was in Study A; that is, fewer ineffective behaviors result in higher scores, whereas on the effective items (1-20) more behaviors result in higher scores. This allows the two sides of the instrument to be summed.

Data were collected on Frame Factor/Context Variables including grade level, SES, and the number of students in various academic strata per class. In addition, the teacher's experience, highest degree, sex, and number of absences over the prior 1.5 years were available. SES measures were school based percentages of students in free lunch programs.

Analyses

In general, analyses paralleled those of Study A. Both Spearman and Pearson correlations were run, from all FPMS total and subscores, against task engagement and two estimates of achievement, and raw posttest scores for each academic group level and each class as a whole.

Since pretest/posttest scores were unavailable, only a covariate, the CAT general reading percentile rank, could be used to define student achievement strata and expected gain scores. Either raw or percentile scores were appropriate for use in the analysis.

ANOVAs were run on FPMS scores to determine possible differences across grades levels.

Results

Study B results are shown for the total scale, made up of 39 quartile-based items, both effective and ineffective. Only the quartile-based scaling is shown, as results in Study A clearly demonstrate the benefits derived from this approach to item and

instrument scoring. This is the scaling used for Master Teacher ranking. Unless otherwise noted, correlations represent FPMS scores against mean posttest class scores for the average stratum. This group, the middle 50 percent on the covariate, provides relatively homogeneous classes across teachers, and tends to moderate various undesirable effects found in simple correlations. It is probably the most stable estimate of the instrument's relationship with student outcomes.

A bivariate outlier was identified in these data, but not included in the results. This is discussed below.

As in Study A, highly significant positive relationships occur between FPMS scores and task engagement measures. In this study, the total score relationship is consistently r=.61 (Table 20).

TABLE 20

Correlations Between FPMS Scores and Task Engagement

	All Students n=39	High Achievers n=38	Average Achievers n=39	Low Achievers n=36
Quartile Total Score	.61	.61	.61	.62
Effective Subscore	.55	.57	.55	.58
Ineffective Subscore	.30	.28	.30	.27

for n = 40
P = .05 at r = .26
P = .01 at r = .36

Unlike Study A, Table 21 depicts zero level correlations between major FPMS scores and achievement as measured by raw posttest and expected gain scores on the American history test.

TABLE 21

Correlations Between FPMS Scores and Achievement

			Average Academic Stratum		
			Task Engagement	Raw Posttest	Expected Gain Score
Quartile		r=	.61	-.04	-.03
Total Score	N=39	P=	.000	.42	.43

In this study, the task engagement measure shows a moderate positive relationship with achievement. The simple correlation between task engagement and raw posttest scores for the average student is r=.263, ($p < .06$).

The relationships found in Study B between major FPMS scores and outcome measures were confounded by an outlier. The data set was therefore, reduced by one case, determined to be extreme in several senses of the term:

1. The individual's FPMS score was 2.5 standard deviations below the sample mean,
2. The individual's task engagement score was over 3 standard deviations below the sample mean,

3. The relationship between the individual's ineffective scores and class posttest scores exhibited extremely unusual characteristics (see plot in Appendix D),
4. The relationship between the individual's class pretest and posttest scores can only be described as anomolous (see plot in Appendix D), and
5. The relationship between the individual's engagement scores and class posttest scores was very unusual. With this individual included in the data, the simple correlation between task engagement and achievement was negative (r=- .014). With this individual eliminated from the data, the relationship between task engagement and achievement became positive and bordered closely on significance (r=.263, P <= .056).

For the above reasons, this person's scores were deleted from the data for these analyses. For more detail on the identification of this outlier, see Appendix D.

The elimination of a single outlier in this study causes the greatest change in the relationship between task engagement and achievement, with the resulting correlation of r = .263 very nearly reaching the generally accepted .05 level of significance (Table 22).

There were also changes in the relationship between FPMS scores and achievement on the order of .14. Although not significant in a single small study, when combined in quantitative meta-analytic analyses, such changes can have substantial effects. The change for task engagement was negligible (-.07).

TABLE 22

Outlier Effects in Correlations

Correlations for Average Students	With Outlier N	Corr	P	Without Outlier N	Corr	P
FPMS Quartile with Task Engagement	40	.68	.00	39	.61	.00
Task Engagement with Achievement	40	-.01	.99	39	.26	.055
FPMS Quartile with Posttest	40	-.18	.14	39	-.04	.42

Analyses between FPMS scores and frame factors and instruction format variables such as grade level and socio-economic status bordered on zero for this study, as they did for Study A ($r = .06$).

The teachers in this study exhibited significantly lower scores on the FPMS summative instrument than those of either Study A, or the FPMS Norming Study. If one were to set cut offs based on the 1985 Florida Master Teacher standards for 75th percentiles, none of the teachers would qualify. Setting 75th percentile rankings from the 1984 FPMS norming study, only one teacher qualifies. Since this could have a significant effect on comparisons between performance scores as they relate to achievement an additional test was run. In this test, teacher scores on the FPMS were classified as being above or below the 50th percentile of the norming study. Specifically the raw posttest score for the average stratum of students for teachers in this study scoring above the 50th percentile from the FPMS norming study were compared to

those scoring below the 50th percentile. Means and the results of Mann-Whitney U tests are shown in Table 23. Just as in Study A, the class mean scores for teachers above the 50th percentile from the norming study are higher than those below, although not significantly higher in this single study. Median scores for top teachers show a greater difference from bottom teachers, as a result of the skewed nature of the distributions.

TABLE 23

Post Elementary Teachers Above and Below the 50th Percentile

Post Elementary Percentile Ranks	Class Mean Raw Posttest Scores for Average Students			
	N	Mean	z	Prob z
Teachers Above 50th Percentile	8	21.36	.71	.48
Teachers Below 50th Percentile	31	20.45		

Summary and Discussion

The American history posttests were distributed to teachers from one to three days prior to administration of the tests. The tests were administered in the usual manner for a semester final exam: (a) in the individual teacher's classroom, and (b) with the teacher serving as proctor. Under these conditions one outlier was found in the data and eliminated when reporting the results. It is recognized that these

conditions do not control sufficiently for teachers' intervention and should be corrected in future studies.

Aside from these concerns, the positive and highly significant relationships between teacher behaviors and task engagement closely parallel those of Study A, and the relationship between task engagement and achievement is positive and borders on significance (.056). In contrast, the simple relationship between the FPMS scores and student achievement, as measured by the American history test, hovers around zero.

The use of a criterion based cut score (50th percentile on the FPMS norming study), does indicate a slightly positive but non-significant effect for higher FPMS scores on student achievement.

Studies A and B were exploratory studies primarily concerned with the question, "Do FPMS scores relate positively to measures of student outcome?"

Instrument Scale Revisions

Data from Studies A and B were submitted to analyses for purposes of instrument scaling.

The conduct of multiple studies using the same independent variable (FPMS summative instrument), allows for extensive investigation of variations in instrument scoring. Two scaling questions were investigated:

1. Is the FPMS quartile-based score an appropriate scaling for this instrument?
2. Are there other scalings, or methods of treating various items on the instrument, that relate either more positively or negatively with student outcome measures than a quartile-based score?

Several score revisions were developed and tested as predictors of student outcomes from Study A that were cross validated with Study B.

These revisions were the following:

1. Revision One - Quartile scaling with an expanded questioning component.
2. Revision Two - Quartile scaling of twenty-three of the instrument's thirty-nine items that best predicted student outcomes in Study A.
3. Revision Three - Scores derived by subtracting the percentage of ineffective scores from the percentage of effective scores.
4. Revision Four - Revision three scores were computed for each item and this score was multiplied by the square root of the absolute number of effective behaviors for the particular item. Scores of each item were then summed to achieve a total score.*

Table 24 depicts the relationships between outcomes for the quartile method of scaling (the scaling used in the Master Teacher Program), and the four scale revisions across the studies conducted to date.

*See Appendix F for additional information on revisions 1-4, the rationale for scaling and the results derived from the study.

TABLE 24

Relationships Between Various FPMS Scalings and Outcomes

			Quar-tile	Rev one	Rev two	Rev three	Rev four
Engagement							
	Study A	n=52	.50	.56	.54	.68	.58
	Study B	n=40	.68	.73	.74	.58	.62
Achievement							
Study A	Math	n=17	.31	.38	.42	.45	.38
	Social Studies	n=30	.28	.39	.37	.38	.27
Study B	History	n=39	-.04	.01	.06	.08	-.17

By reference to Appendix F, it can be seen that the quartile method of scaling correlated r = .90 with all revisions with the exception of revision three where r = .74. These data show that revisions one and three relate more strongly with student outcomes than quartile scaling, however, until further tests of scaling are conducted, the current method of scoring the instrument appears adequate. Further cross validation of scaling will occur in Studies C, D, and E.

Data from Studies A and B and the scaling tests provide a basis for further investigation of FPMS summative instrument scores in relation to the combination of several outcomes.

Meta-Analytic Approach to Multiple Studies

The quantitative approach to multiple studies provided by meta-analytic techniques allows the combination of several outcomes, thereby increasing the power of test statistics. Maximizing power is particularly useful for studies having small numbers of cases (as is typical in studies of teacher effectiveness), or showing small but consistent effect.

Rosenthal (1978) suggests the use of a combined Z based upon the P values of specific relationships to test the overall significance of results. The combined Z is computed using the following formula:

$$Z = \sum z / \sqrt{N}$$

where z = the z associated with a specific one-tailed probability, and

N = the number of studies

Since the first two studies contained several subsets of independent variables (of particular interest are differences in relationships for different subject matters or grade levels), the following discussions concern several combined Z's. Most of these are represented by only one or two studies. As studies continue on this topic, the power of combined analyses to identify important but moderate effects will increase. It is important when combining studies to use only combined Z's when the number of studies is small, on the order of three to five. When the results of many studies are combined, such significance tests are no longer appropriate.

Combining Correlation Coefficients

The following section displays combined results of the two studies conducted to date. Table 25 shows combined Z's for the relationship between FPMS quartile scores and task engagement (note: this relationship is of sufficient strength that combined Z's are superfluous, as all three sets of relationships are independently significant at the .001 level).

TABLE 25

Combined Z Scores Between FPMS Scores and Task Engagement

Category	N of Cases	Task Engagement	
		Comb Z	P
All Studies	3	5.38	.0001
Elementary	2	4.04	.0001
Post-Elementary	1	3.60	.0001
Math	1	2.75	.0001
Soc Std/Hist	2	4.63	.0001

Table 26 depicts combined Z's resulting from the direct relationships between two FPMS scores (the quartile scaling and one scale revision), and raw posttest scores for average students (this stratum is defined as the average stratum on pretest scores). This table indicates a significant relationship between FPMS scores and student achievement for the quartile scaling only in the elementary grades, although the results

of all studies closely approach significance (.067). Revision One shows a significant relationship across all studies (.02), a highly significant relationship for elementary (P <= .01), and a relationship approaching significance for math and non-math classes at the secondary level (.07).

TABLE 26

Combined Z Scores between FPMS and Achievement

Category	N of Studies	Raw Posttest Scores			
		Quartile Scores		Revision One	
		Comb Z	P	Comb Z	P
All Studies	3	1.50	.067	2.13	.017
Elementary	2	1.96	.025	2.60	.005
Post-Elementary	1	- .17	- .433	.02	.492
Math	1	1.22	.109	1.50	.067
Soc Std/Hist	2	.98	.116	1.55	.061

The relationship between task engagement and achievement is positive for all three study areas, significant (.05) for the elementary social studies classes, and nearly significant in the American history class (.06). The combined z scores for all groups are highly significant (.01) for all studies and for non-math studies, and significant (.05) for elementary studies (Table 27).

TABLE 27

Combined Z Scores between Task Engagement and Achievement

Category	N of Cases	Task Engagement	
		Comb Z	P
All Studies	3	2.37	.01
Elementary	2	1.77	.04
Post-Elementary	1	1.60	.06
Math	1	.70	.25
Soc Std/Hist	2	2.41	.01

A Criterion Approach to the Analysis of Predictive Studies

To qualify for the Florida Master Teacher Program, teachers must receive a superior classroom performance score. Superior performance has been defined as a score at or above the 75th percentile of all applicants for Associate Master Teacher Status. Such a standard may prove too rigorous in the research environment, since only 25 percent of the teachers by definition can reach such a score. This seems artificially high given the number of tests used to select Associate Master Teachers. For purposes of this research, the 50th percentile of the FPMS norming study was used to divide teachers into top and bottom groups. Using such a division, one can measure differences in student outcomes between teachers scoring in the upper one-half of the performance score distribution and those scoring in the lower one-half.

Of the 82 teachers taking part in Studies A and B, only four (five percent) would have qualified under the Associate Master Criteria of the 75th percentile. Thirty-two of 81 (39 percent) of the teachers in Studies A and B scored at or above the 50th percentile of the norming study. It is therefore assumed that the studies conducted to date for purpose of predictive validity may not be representative of teachers throughout the state.

Tables 28 and 29 show that the average students in top teachers' classes scored higher (although not significantly higher) than average students in bottom teachers' classes in both studies. The results of such comparisons are somewhat clouded by the teachers in the middle of the distribution, who, although scoring very nearly the same on the FPMS instrument, are arbitrarily assigned either a top or bottom designation. For example, significance tests run comparing average task engagement for top groups against bottom groups proved non-significant despite the sizable and significant correlation between this variable and FPMS scores.

The teachers in Study B exhibited lower scores on the FPMS summative instrument than those of either Study A, or the FPMS Norming Study. If one were to set cut offs based on the 1985 standards for 75th percentiles, none of the teachers would qualify. Setting 75th percentile rankings from the 1984 FPMS norming study, one teacher qualifies.

The percentile ranks achieved by classes of teachers scoring above and below the 50th percentile are shown in Table 30. In all cases, the

TABLE 28

Elementary Teachers Above and Below the 50th Percentile

Elementary Percentile Ranks	Class Mean Raw Posttest Scores Average Stratum			
	N	Mean	z	Prob z
Teachers Above 50th Percentile	14	17.26	1.53	.13
Teachers Below 50th Percentile	32	16.05		

TABLE 29

Post Elementary Teachers Above and Below the 50th Percentile

Post Elementary Percentile Ranks	Class Mean Raw Posttest Scores for Middle 50% of all Students			
	N	Mean	z	Prob z
Teachers Above 50th Percentile	8	21.36	.71	.48
Teachers Below 50th Percentile	31	20.45		

pretest percentile ranks were 50. As the table shows, for math, social studies and American history, teachers scoring above the 50th percentile show students with improved posttest percentile ranks, while those below the 50th percentile show average students with reduced percentile ranks.

TABLE 30

Average Posttest Percentile Ranks

	Study A		Study B
	Math	Social Studies	American History
	Percentile Ranks of Average Stratum		
Teachers at or above 50th percentile	73	55	57
Teachers below 50th percentile	45	45	46

Combining the Z scores obtained in the above analyses results in a z of 1.58, and a respective probability level of .056. Thus we may be about 94 percent sure that differences on achievement tests exist between teachers scoring above and below the 50th percentile of the Norming Study.

These results provide supporting evidence for the use of FPMS in the Master Teacher Program, however, serious consideration should be given to lowering the standard from the 75th percentile to the 50th percentile for these purposes.

Conclusions Based on Meta-Analytic Techniques

The preceding analyses indicate:

1. There is a highly significant relationship between FPMS scores and task engagement,
2. There exists a highly significant relationship between task engagement and student achievement,
3. There exists a significant relationship between FPMS scores and student achievement, and
4. Teachers scoring at or above the FPMS norming groups 50th percentile probably have more positive effects on their students' achievement than those scoring below this point (at .06 level of confidence).

These results suggest that there is a direct relationship between FPMS scores and student behavior, that there is a direct relationship between student behavior and achievement, and that the indirect relationship between teacher behavior and student achievement results partly from activities assigned by teachers.

Chapter V

Development of Student Services Personnel Observation System

Certain individual and group services, primarily consultative, are concerned directly or indirectly with various student needs such as counseling, personal assistance, or physical welfare. We refer to individuals who render these services as student services personnel. These individuals are not, strictly speaking, teachers. In the performance of their functions, they do not provide instruction in classroom, laboratory, shop, or playground. Neither are they staff officers; they do not function as assistants or advisors to the administrative personnel, as do supervisors or other staff officers.

Among the student services personnel are the following: occupational placement specialists, school counselors, school psychologists, visiting teachers, school social workers, and student services administrators. Although not teachers, they are designated by the state as instructional personnel. As such, they come under the Florida Master Teacher Program and are entitled to apply for status in that program.

Since student services personnel do not function as teachers, their performance on the job may not be adequately measured by the FPMS summative instrument. Consequently, a new observation system was designed to code their on-the-job behavior. Steps in the development of this system are set forth in the following pages.

Development of Instrument Content

The first step in the development of an observation system is to select the behaviors to be observed. This step can be taken in either of two ways or by a combination of the two. The behaviors can be determined by the judgment of knowledgeable individuals or by recourse to research literature. The first of these is based on professional lore and is subject to the accumulated errors of crude experience, but in the absence of research, such lore is not to be ignored. The second approach entails a survey of relevant research. A combination of these approaches was followed in the selection of the concepts and indicators comprising the content of the observation system for student services personnel.

A task force consisting of district supervisors and student services personnel was formed to review the research literature on student services. From research and professional consensus, the task force compiled a set of six concepts and their performance indicators for student services personnel. The concepts thus derived are as follows: management of counseling sessions, development of rapport, communication skills, interpersonal skills, problem/concern clarification, and development of plans. Forty-eight indicators for these concepts were formulated, together with examples. The team also developed eight concepts for consultants, as follows: management of counseling sessions, agenda building, problem definition, interpersonal skills, communication skills, information skills, development of plans, and termination of consultation session. Fifty-one indicators plus examples were formulated for these concepts.

Validation of Content

Both sets of concepts and indicators were submitted to approximately 900 student services personnel, who were asked for their opinion on the appropriateness and comprehensiveness of these items. Four hundred sixty-four individuals (52 percent) responded to the request. The distribution of these respondents by position is shown in Table 31. Almost half of the respondents are counselors, a proportion not out of line with the proportion in the state at large.

TABLE 31

Distribution of Respondents by Positions

Position	Frequency	Percent
Occupation Placement	39	8.4
Counselor	212	45.7
Psychologist	91	19.6
Visit/Social Work	39	8.4
Services Administration	59	12.7
Other	24	5.2
TOTAL	464	100.0

The distribution by years of experience can be seen in Table 32. A little over 83 percent have four or more years of experience, a fact of some importance since three years is typically considered the minimum of experience for job adjustment for beginning school personnel.

TABLE 32

Distribution of Respondents by Years of Experience

Experience	Frequency	Percent
1 year or less	14	3.0
1 to 3 years	62	13.4
4 to 9 years	145	31.3
10 to 14 years	133	28.6
15 or more years	110	23.7
TOTAL	464	100.0

According to the distribution of respondents by school level (Table 33), 28 percent work in elementary schools, 11.6 percent in junior-middle schools, 16.4 percent in senior high schools, and 30.4 percent in some combination of these levels. Respondents working in vocational schools and other offices comprised 13.6 percent of the sample. These data suggest a reasonable representation of personnel across position, experience, and grade level.

It can be seen from Table 34 and Table 35 that the concepts for counseling and consulting respectively evoked almost complete agreement on their appropriateness among respondents. For counseling, the concepts least favored were development of plans and problem/concern clarification, each of which was rejected by only 3.7 percent of the respondents. For consulting, problem definition and interpersonal skills were rejected by 4.3 and 3.1 percent respectively.

BEST COPY AVAILABLE

TABLE 33

Distribution of Respondents by School Level

School Level	Frequency	Percent
Elementary	130	28.0
Junior-Middle	54	11.6
Senior	76	16.4
Combination	141	30.4
Vocational	21	4.5
Other	42	9.1
TOTAL	464	100.0

The results of the literature search, the consensus of the review team, and the judgment of 464 individuals active in the field support the content validity of this domain.

TABLE 34

Opinions of Concept Appropriateness-Counseling

Concepts	Percent Yes	Percent No	Percent Undecided	Missing	Total
Management of Counseling Sessions	96.3	1.9	1.8	0.0	100.0
Development of Rapport	99.4	0.4	0.2	0.0	100.0
Communication Skills	95.7	2.2	2.2	0.0	100.0
Interpersonal Skills	91.8	4.3	3.9	0.0	100.0
Problem/Concern Clarification	89.2	3.7	7.1	0.0	100.0
Development of Plans	89.2	3.7	7.2	0.0	100.0

TABLE 35

Opinions of Concept Appropriateness-Consulting

Concepts	Percent Yes	Percent No	Percent Undecided	Missing	Total
Management of Counseling Session	94.8	2.0	3.3	0.0	
Agenda Building	96.3	1.3	2.4	0.0	
Problem Definition	92.6	4.3	3.0	0.0	
Interpersonal Skills	94.8	3.1	2.2	0.0	
Communication Skills	96.5	1.5	1.9	0.0	
Information Skills	96.9	0.9	2.2	0.0	100.0
Development of Plans	97.4	1.1	1.5	0.0	
Termination of Consultation Session	95.7	1.5	2.8	0.0	

Although two sets of concepts, one for counseling and one for consulting, were selected, similarity among them led to the development of a single observation instrument consisting of 19 effective and 18 ineffective indicators. This instrument was then subjected to the scrutiny of a review team for examination of item clarity and topography. (Appendix G contains a copy of the instrument.)

Estimation of Instrument Reliability

General conditions for estimating the reliability of this instrument are: (1) the consistency with which a score on the instrument differentiates among a group of subjects, (2) the consistency with which different coders record the same behaviors in a similar fashion for different subjects, and (3) the consistency of a subject's score over time or situations.

Several approaches to obtaining these estimates are used at the present time, including:

1. the generalizability (intraclass) approach,
2. the correlation approach, and
3. intercoder agreement estimates.

One of the strongest of these approaches is that of generalizability. This form of analysis was applied to the data obtained in this study. In addition, coefficient alpha, an estimate of internal consistency was computed.

A sample of 100 observations was obtained, including 28 pairs of observations conducted at two different times by two different observers on the same subject.

The sample consisted primarily of white (81 percent) females (67 percent) holding a master's or higher degree (90 percent). Most observations took place in an urban setting (97 percent), with 24 percent of the observations at the elementary level, 23 percent middle, six percent secondary and 46 percent across more than one level. The median length of observation was 20 minutes (mean 25 minutes). Observation length ranged from 15 to 60 minutes, with 43 (43 percent) taking only 15 minutes.

Forty-nine observations of counselors were made, four of occupational therapists, 21 of psychologists, and 12 of social workers. Session types included 45 counseling, 29 consulting, and 11 defined as some other type. Thirty-eight of the sessions were with individuals, while 40 were in a group situation. The number of participants in these

sessions ranged from one to 64, with a median group size of 4.7. (The mean of 11.9 was pulled far out by a few extremely large groups.) Thirty-eight of these sessions occurred with students, 16 with faculty members, six with parents alone, four with parents and faculty, and seven with various combinations of individuals.

Much as with the FPMS summative instrument, considerable variation occurred for the effective side of the instrument, with little or no variation for the ineffective behaviors. Effective total scores ranged from two to 19, with a mean of 12.4 (median 13.1). Ineffective total scores ranged from 10 to 18, with a mean of 16.5 (median 17.0). Combined scores ranged from 10 to 37, with a mean of 28.9 (median 29.4).

TABLE 36

Three Reliability Estimates for Student Services Instrument

	Various Reliability Estimates		
	Intraclass Discriminant Reliability n = 28	Coefficient Alpha n = 100	Intraclass Intercoder Agreement n = 28
Complete Total Score	r = .58	r = .78	r = .56
Total Effective Score	r = .59	r = .57	r = .37
Total Ineffective Score	r = .24	r = .23	r = .20

*It should be noted that Intraclass estimates for the study design used are extremely conservative.

As Table 36 shows, the highest reliability estimates result from the total score, derived by summing effective and ineffective scores on the instrument. The scoring of ineffective items is reversed, so that fewer ineffective behaviors result in higher scores, thus allowing the two sides to be summed. Few behaviors occurred on the ineffective side, which resulted in considerably reduced variance and low reliability estimates. Therefore, total score estimates should be used in any evaluation conducted using this instrument.

These tests suggest minimum reliability of the instrument when applied to individual and group consulting and counseling situations.

Materials produced for training observers in the student services evaluation system include:

1. Domain Documents (Domain 7 and 8 of the Florida Performance Measurement System)
2. Coding Manual
3. Observation Instrument
4. Videotapes

The observation instrument is included in this report as Appendix G. The domain documents and coding manual may be obtained from Dr. Garfield Wilson, Office of Teacher Education, Certification, Staff Development and Professional Practice Services, Department of Education, Tallahassee, Florida.

Recommendations

1. Some concepts and indicators in these domains which are currently based solely upon consensus should be submitted to empirical tests to establish their effects on client populations.
2. Further tests should be conducted to verify instrument reliability, generalizability, and practicality during actual application. Data collected on observations for the Master Teacher Program should serve towards this end.

Chapter VI

Development of Data Collection System for Library Media Specialists

Like student services personnel, library and media specialists are not, strictly speaking, teachers. Their primary function is to collect, organize, and make available learning materials for students and to provide instruction in such activities as locating and using these materials. They are not classroom teachers. Neither are they line nor staff officers as are principals, supervisors, and assistant superintendents. They are designated by the state as instructional personnel and therefore entitled to apply for status in the Florida Master Teacher Program.

Since the library media personnel provide some instruction as noted above, their performance in this regard can be adequately measured by the summative instrument of FPMS. A substantial proportion of their work is not covered by this instrument, however; and consequently an additional instrument was designed to evaluate their on-the-job behavior and products for this part of their work. Steps in the development of this instrument are set forth under the following headings:

1. Development of instrument content.
2. Validation of content.
3. Estimation of instrument reliability.

Development of Instrument Content

As in other pedagogical instruments, the first step in the development of an instrument is to select the items to be evaluated. This step can be taken in either of two ways, or by combination of the two. The items can be determined by the judgment of knowledgeable individuals or by recourse to research literature. The first of these is based on professional lore and is subject to the accumulated errors of crude experience. The second approach entails a survey of research literature. A combination of these approaches was used in the selection of the concepts and indicators comprising the content of the evaluation instrument for library media specialists.

Library and media specialists were retained to search the literature for evidence of effective and ineffective practice in these fields. A task force consisting of library and media specialists from the various school settings, including vocational education, was then assembled to examine the concepts and indicators compiled from research and to determine by consensus additional concepts and indicators.

The work of the task force resulted in a set of nine concepts: planning, organization, evaluation, selection/acquisition, maintenance, skills instruction, consultation/information, production, and public relations. Thirty-two indicators for these concepts were formulated, together with examples. These concepts, indicators, and examples together with supporting evidence from research, were organized as Domain 9 of the Florida Performance Measurement System.

Validation of Content

These concepts and indicators were submitted to approximately 900 library media specialists, who were asked for their opinions on the appropriateness and comprehensiveness of the material. Seven hundred and twenty-seven individuals responded to the request. Distribution of these respondents by years of experience is shown in Table 37. As can be readily seen, approximately 80 percent of the respondents had five or more years of experience.

TABLE 37

Distribution of Respondents By Years of Experience

Years	Frequency	Percent
1 year or less	39	5.4
2 to 4 years	104	14.3
5 to 9 years	160	22.0
10 to 14 years	196	27.0
15 or more years	228	31.4
TOTAL	727	100.0

It is interesting to note that in addition to this heavy loading of experience, 77 percent of the respondents held an in-field degree, with a little over 56 percent holding the master's degree (Table 38).

It is important to know the school level these respondents represent, for it is conceivable that the job requirements can be influenced by school level. As can be seen from Table 39, approximately 53 percent

TABLE 38

Distribution of Respondents By Levels of Training

Degree	Frequency	Percent
Bachelor	148	20.4
Master	374	51.4
Advanced Graduate	38	5.2
No Degree/Certified	152	20.9
Not Certified in Media	12	1.7
No Response	3	0.4
TOTAL	727	100.0

were from elementary schools and about 41 percent from secondary schools.

TABLE 39

Distribution of Respondents By School Level

Levels	Frequency	Percent
Elementary	387	53.2
Junior-Middle	138	19.0
Senior	162	22.3
Vocational	16	2.2
Other	12	1.7
No Response	12	1.7
TOTAL	727	100.0

The vocational schools were represented by a little over two percent. In view of the relative proportion of vocational schools to elementary and secondary schools in Florida, this percentage does not seem to be out of proportion.

Another factor that may influence the job requirements is the school size. It is therefore interesting to note that approximately 69 percent of the respondents were from schools with enrollment of from 500 to 2000 students (Table 40). Approximately one-fifth of the respondents came from schools with enrollments from 250 to 500 students and slightly more than six percent came from schools of 2000 or more.

TABLE 40

Distribution of Respondents By Size of School

Enrollment	Frequency	Percent
Up to 249	23	3.2
250 to 499	154	21.2
500 to 999	315	43.3
1000 to 1999	185	25.4
2000 or more	49	6.7
No Response	1	0.1
TOTAL	727	100.0

These tables suggest a reasonable representation of personnel across experience, grade level, school size, and levels of training.

It can be seen from Table 41 that the respondents agreed almost completely that the nine concepts were appropriate. The least

acceptable concept, production, received a little over 81 percent positive responses. The library media specialists evidently do not consider the design and preparation of audio visual materials to be as important as other duties. The next least acceptable concept was evaluation, which requires the library media specialist to gather and analyze data about the use of the library and media services and resources. Almost 93 percent consider this responsibility to be important. The remainder of the concepts were favorably accepted by some 95 to 100 percent of the respondents.

TABLE 41

Distribution of Opinions of Concept Appropriateness

Concept	Yes	No	Undecided	Missing	Total
Planning	98.5	1.0	0.6	0.0	100.0
Organization	98.9	0.6	0.6	0.0	100.0
Evaluation	92.6	2.5	4.9	0.0	100.0
Selection/Acquisition	99.7	0.1	0.1	0.0	100.0
Maintenance	98.9	0.3	0.8	0.0	100.0
Skills Instruction	94.5	1.8	3.7	0.0	100.0
Consultation/Information	100.0	0.0	0.0	0.0	100.0
Production	81.8	8.7	9.5	0.0	100.0
Public Relations	95.8	0.8	3.2	0.0	100.0

The compilation from the literature search and the judgment of 727 respondents, active in library media work, support the content validity of this domain.

The nine concepts were classified into three functions -- management, collection-development, and services. Planning, organization, and

evaluation constitute the management function; selection/acquisition and maintenance make up the collection-development function; and skills instruction, consultation/information, production, and public relations constitute the service function. These three functions were used to organize the nine concepts and 25 of the 32 indicators into a data collection instrument for library media specialists. (For the instrument thus developed see Appendix H.)

It should be noted that this instrument differs from other summative instruments of the FPMS in the sense that no ineffective indicators are included. The library media specialist either performs an indicator or does not, and the instrument provides for the performance of the indicator to be recorded and nonperformance likewise to be recorded.

Estimates of Instrument Reliability

Factors to be considered for claiming the reliability of this instrument are: (1) the consistency with which a score on the instrument differentiates among a group of subjects, (2) the consistency with which different coders record the same behaviors in a similar fashion for different subjects, and (3) the consistency of a subject's score over time and/or situations.

Several approaches to obtaining these estimates are commonly used at the present time, including: (1) the generalizability (intraclass) approach, (2) the correlation approach, and (3) intercoder agreement estimates. One of the strongest of these approaches is that of generalizability. This form of analysis was applied to the data

obtained in this study. In addition, an estimate of internal consistency, coefficient alpha, was computed.

A sample of 25 pairs of ratings for various library media specialists was obtained. The scores reported result from the average of a pair of coders for a given subject. The sample was 92 percent female, 88 percent white, and 67 percent master's degree holders. The mean experience was 11.7 years and ranged from two to 25 years. Ninety-two percent of the ratings took place in urban settings. Specialists were spread over grade levels with ten at elementary, eight at middle schools, and seven at the secondary level. The staffing of these centers varied, although each had at least one or two additional staff members. The largest staff consisted of six individuals. The number of professional personnel ranged from zero to three, with 57 percent having one or less; most settings had at least one nonprofessional staff member. Fully 85 percent of these specialists claimed multiple responsibilities, including audiovisual, television and print services.

As Table 42 indicates, the reliability estimates for this instrument are relatively high for the total instrument (r=.77 to r=.89, high for the management subsection (r=.71 to r=.81), and low to moderate for the collection/development and service sections (r = .63 to r = .70). Therefore, total score estimates should be used in any evaluation conducted with this instrument.

Intraclass reliability estimates are very high, given the nature of the data, suggesting this is a very reliable instrument.

TABLE 42

Three Reliability Estimates for Library Media Instrument

	Number of Items	Intraclass Discriminant Reliability 25 pairs	Coefficient Alpha n = 25	Intraclass Intercoder Agreement n = 25
Total Score	25	r = .77	r = .89	r = .79
Management	11	r = .71	r = .81	r = .73
Collection Development	5	r = .66	r = .70	r = .69
Services	9	r = .63	r = .72	r = .64

Materials produced for training persons to code the library media instrument include:

1. Domain document (Domain 9 of the Florida Performance Measurement System)
2. Coding manual
3. Data collection instrument

The instrument is included in this report as Appendix H. The domain documents and coding manuals may be obtained from Dr. Garfield Wilson, Office of Teacher Education, Certification, and Staff Development, Department of Education, Tallahassee, Florida.

Recommendations

1. Some concepts and indicators in this domain which are currently based solely upon consensus should be submitted to empirical tests to establish their effects on client populations.

2. Further tests should be conducted to verify instrument reliability, generalizability and practicality during actual application. Data collected by coding this instrument for the Master Teacher Program should serve towards this end.

Chapter VII

Development of Addendum to the Coding Manual for Exceptional Student Education Specialists

The studies of the Florida Performance Measurement System that were conducted during 1983-1984 and 1984-1985 supported the generic nature of the concepts and indicators to a greater extent than was expected. Similarities in teaching were demonstrated by all teachers with their use of questioning, reinforcing, providing for practice, giving feedback, stopping misconduct, and numerous other common skills. Not only did all effective teachers use these skills, but they used them appropriately. For example, stopping misconduct needs to be done promptly, on target, and without being punitive or sarcastic. Data from the norming studies indicated that differences in the classroom performance of effective teachers did not vary significantly except between instructional levels.

All teachers exhibited the teaching behaviors listed on the summative instrument, but there was a difference in the use of behavioral indicators between K-5 teachers and 6-12 teachers, thereby establishing two groups. All other factors, such as subject areas, were not significantly different.

Teachers of different subject areas with students of varying academic talent and interests found it difficult to accept one set of criteria for evaluating their performance. This was particularly true of teachers of exceptional students, sometimes referred to as special education teachers, whose students have mild to severe mental and physical impairments.

Although this group of teachers was a part of the norming study population, they suggested that the domains and the behavioral indicators on the summative observation instrument were not quite appropriate for them because few of the studies described in the domains included exceptional students. They proposed a study to determine if the teacher behavior in the FPMS was adequate to describe the classroom performance of special education teachers. Therefore, a task force of specialists was selected from the various areas of special education to further examine the research base and instrument for applicability to this area.

The Exceptional Student Education (ESE) Task Force searched the literature for studies dealing with exceptional students. From the research, they developed special competencies in the various areas of exceptionality. Practitioners rated the competencies indicating the extent to which the skill or knowledge represented in the competency was needed by the special area teacher. In a review of this consensus study, the concepts and indicators represented by these competencies were found to be special cases of the generic concepts and indicators on the FPMS observation instrument.

The task force decided, as a result of their analysis, to use data from the consensus study to write examples from special education in a coding manual that could be used in training observers when using the FPMS summative instrument for coding teachers of special education classes. The task force reviewed each concept and its indicators, as set forth in the regular coding manual. Wherever examples were needed

to indicate the applicability of descriptors to the unique requirements of teacher performance in exceptional education, appropriate examples were added.

A draft of an Exceptional Student Education Coding Manual instrument was developed and submitted to 225 practitioners for content validation. Data from 125 respondents were analyzed by our staff for the revision of the ESE coding manual. (See Appendix I for the report on the revised coding manual). In short, specialists in the area of Exceptional Student Education retained the FPMS summative instrument for observations of their practitioners, and modified the coding manual to reflect more exactly how behaviors on the instrument would appear in the exceptional education classroom. Data are being monitored to determine effects on reliability and norming distributions.

Recommendations

1. Further tests should be conducted with increased subsample sizes of the Norming Study, by specialized areas within special education, to verify the reliability and generalizability of the FPMS summative instrument in these areas. The Master Teacher data could serve this end.
2. The effects of providing a separate coding manual for observers of special education teachers only should be studied. The additional behaviors listed in this manual as effective may give special educators an unfair advantage in competing for Master Teacher status or in fulfulling Beginning Teacher requirements.

The fact that a disproportionate number of Special Educators qualified for Associate Master Teacher status (36 percent), calls into question the practice of creating separate or additional manuals.

Chapter VIII

Master Teacher Program:

System for Assigning Additional Evaluation Points

The legislation establishing the Master Teacher Program specifies the following conditions for participation in the program at the Associate Master's level:

1. Four years teaching experience, at least two of which must have been in Florida.
2. A professional service contract or a continuing contract.
3. A masters degree-in field from a standard institution. If the masters degree is out-of-field, then 15 semester hours in-field at the graduate level must be completed.
4. Successful performance on an academic examination in-field.
5. Successful classroom performance as determined by a valid, reliable, and normed observation instrument approved by the State Board of Education.

In addition, the legislation provided for additional points, as follows: "If after conducting the evaluation of an applicant pursuant to the performance evaluation system the principal determines that the system does not adequately recognize the performance of the applicant, the principal may upon stated reasons, recommend additional evaluation points for the applicant. In such event, the evaluation shall be subject to review and approval by a third party, successfully trained in the use of the performance evaluation system and approved by the Commissioner of Education."

This additional condition presents a number of perplexing questions, such as: What teacher behaviors are not covered by the observation instrument? What kind of supporting evidence is required of the principal? How can the data be collected and processed? How can the points given by principals be accorded comparable treatment?

The first question is how to determine the extra teaching behaviors to be considered by principals. This question is answerable in either of two ways. The most obvious way is to leave the scope of teaching behaviors entirely open, relegating to the principal the responsibility for deciding what behaviors are not included in the performance evaluation system. On its face, administration of this procedure appears to be simple. However, a bit of reflection indicates that the various interpretations of behaviors among some 4000 principals would present an overwhelming task of data classification and processing. Variation in the selection of additional points for thousands of teachers would require a complex system of review to determine whether or not the behaviors of the evaluation system were duplicated and whether or not the behaviors selected by the principals were in fact teaching behaviors.

An alternative approach is to suggest a set of teaching behaviors not included in the observation system within which the principal can exercise maximum freedom of choice, as long as the choices are consistent with the requirements of a reasonable and efficient scheme of data collection and processing.

Content Identification

We chose the latter approach, which required that our first task be that of surveying the possible sources of concepts and principles not included in either FPMS or TADS. Our search led us to three primary sources:

1. Research on teacher effectiveness. This research has been assembled, organized, and issued under the title Domains: Knowledge Base of the Florida Performance Measurement System. All of the concepts and indicators found in this monograph were compiled into a list against which the concepts and indicators of FPMS and TADS were checked. Teaching performance not found in these systems were then listed as candidates for the checklist for awarding additional points.
2. Evaluation instruments of Florida school districts. Our previous study of these instruments (Teacher Evaluation Project, Final Report for 1982-1983, Coalition for the Development of a Florida Performance Measurement System) provided us with a frequency compilation and distribution of items across two sets of categories: those having to do with classroom instruction and those related to non-classroom responsibilities. From this analysis, we then developed a list of classroom behaviors found in the instruments developed and used by a sample of Florida school districts.
3. Florida Essential Competencies. These competencies were analyzed to determine those not covered by either FPMS or TADS.

From these three sources, 19 concepts and their indicators were compiled, defined, and organized into a checklist (Appendix J). These were subsequently reviewed by four knowledgeable persons for clarity, comprehensiveness, and organization.

Consensus Validation

Although this list of concepts and indicators was derived from research and from sources that had previously been subjected to consensus validation, it was deemed advisable to subject the list to still another consensus study. Since the concepts and indicators were to be used by principals to meet the legal requirements for additional merit points, it seemed appropriate that a sample of principals review the list and rate the concepts as to their relevance and inclusiveness, and to add such additional concepts as they deemed appropriate.

The total list of 19 concepts and indicators was submitted to the principals of three school districts -- Broward, Pasco, and Sarasota -- for review and rating of concept importance. (For instructions for rating check-list categories, see Appendix J.) A total of 106 principals from Broward, Pasco, and Sarasota districts reviewed the list and rated the concepts. It can be seen from Table 43 that a little more than a third of the sample is from the Broward district, almost half from the Pasco district and a fifth from the Sarasota district.

The sample represents schools of various sizes and levels. Table 44 shows that 25 percent are principals of high schools, four percent junior high schools, 13 percent middle schools, 46 percent elementary schools, and 11 percent unspecified.

TABLE 43

Distribution of Principals by Districts

Name	Number	Percent
Broward	37	35
Pasco	48	45
Sarasota	21	20
TOTAL	106	100

TABLE 44

Distribution of Principals by School Levels

Name	Number	Percent
High School	27	25
Junior High	4	4
Middle School	14	13
Elementary	49	46
Unspecified	12	11
TOTAL	106	99

The schools range in size (Table 45), as measured by the number of teachers from 25 or less to 201 or more, with the major portion of the sample representing schools of 26 to 100 teachers. These data indicate that a substantial number of responses to the additional point checklist were received from all three districts, that all school levels were represented, and that small to large schools were sampled.

TABLE 45

Distribution of School Size by Number of Teachers

Number per School	Number	Percent
0 to 25	8	7
26 to 50	56	53
51 to 100	30	28
101 to 200	8	7
Over 201	1	1
Unspecified	3	3
TOTAL	106	99

The ratings by the 106 principals of the 19 concepts comprising the additional-point list were rank-ordered by group-mean rating from x=5.70 for Professional Conduct to low x=2.71 for InterCultural Education (Table 46). Categories of group-mean scores below 3.07 were deleted, leaving a total of 15 items in the check list. These items subsequently became the additional-point checklist used by principals throughout all school districts to assign additional points in cases where the principal determined that the FPMS or TADS (in Dade district) did not adequately cover the performance of the applicant for associate master teacher status.

TABLE 46

Content Validation Data Summary Additional-Point Checklist

Items Ranked High to Low Ratings by Total Sample of 106 Principals.

Item	Total Group Mean	Total Group Rank
15. Professional Conduct	5.70	1
6. Evaluation	5.63	2
16. Professional Growth	5.33	3
11. Planning for Instruction	3.89	4
1. Knowledge of Subject Matter	3.81	5
2. Knowledge of Language	3.68	6
9. Teacher Expectation	3.62	7
8. Interpersonal Relations with Students	3.51	8
12. Individualization	3.49	9
14. Innovativeness and Independence	3.47	10
17. Counsels with Students	3.45	11
3. Knowledge of Computational Skills	3.42	12
7. Workroom Atmosphere	3.39	13
10. Physical Environment	3.30	14
13. Uses Multi-Media	3.07	15
18. Peer Collaboration	2.98	16
4. Fosters Intellectual Skills	2.87	17
19. Management	2.79	18
5. Intercultural Education	2.71	19

Scoring Procedure

The Department of Education interpreted the intent of the legislation creating additional-point assignment by the principal as giving maximum authority to the principal in deciding how these points were to be assigned to Master Teacher Program candidates. Therefore, the Department ruled that ten points could be assigned across all fifteen concepts on the instrument; or all ten points could be assigned to one concept; or up to ten points could be distributed in any proportion the principal felt appropriate.

The ten points, or any proportion thereof, were then scaled to each instrument, providing a maximum of 10 percent of the median score from the norming of the instrument (FPMS 3.3, TADS 6.0).

Recommendations

1. Attaining equality of evaluation among candidates, a basic tenent of teacher evaluation, is compromised when principals can assign ten points to one category while others spread these points among a larger number of categories -- up to fifteen. This procedure should be carefully examined and alternative procedures tested to better standardize the value of these points across candidates.
2. Distributions of performance scores on each instrument (i.e. FPMS, TADS, Library Media and Student Services) from the 1985 Master Teacher Program observations should be examined to determine the power and effects of additional evaluation points. THIS IS A CRITICAL RECOMMENDATION since variations in shapes of distributions will cause the same additional points to qualify a substantially larger proportion of candidates under one system than the other. It also raises the question of validity of a system producing distributions with extreme skewness, and a mean score only three points below the maximum possible score.

Chapter IX

A Comparative Study: FPMS with TADS

State Board of Education Rule 6A-4.45(5)(e)2. mandates that each performance measurement system used in the Master Teacher Program meet specified criteria as judged by a three member review team. Accordingly, two systems were approved -- the Florida Performance Measurement System (FPMS) and the Teacher Assessment and Development System (TADS). (Appendices K and L contain the review reports for FPMS and TADS.

Since two distinct systems were used to measure performance of essentially the same group of people (Master Teacher Program applicants), the state was concerned that, regardless of which system was used, resulting selections of persons to receive awards be equitable across the State. The purpose of this research was, therefore, to develop an information base to which three major questions may be addressed:

1. Do FPMS and TADS assess similar dimensions of teacher behavior as determined by: domain content, and item and definition similarity?
2. Do FPMS and TADS rank scores correlate and do they order teachers in a like manner on a scale of effectiveness?
3. How do FPMS and TADS scores relate to various measures of student achievement?

A plan to address the first question was developed under this contract. Specifically, this plan includes the following:

Cross-validation of FPMS and TADS items.

1. Assemble a team of knowledgeable persons: two knowledgeable of FPMS, two knowledgeable of TADS, and one outside expert on teacher effectiveness literature.
2. Compare the research base and organization of the domains and compare the definitions and examples for each item to determine similarities and differences between instruments.
3. Construct three lists of items.
   a) Items judged equal on both instruments
   b) Items peculiar to FPMS
   c) Items peculiar to TADS
4. Prepare a written analysis of these findings.

This study will require approximately three days of comparative analyses and two days to analyze and summarize the findings. It is to be conducted in the fall of 1985.

During May of 1985, a study was conducted that addresses the second research question. A report of that study follows.

Design

This question can best be answered by relating scores from FPMS and TADS simultaneously drawn from a sample of practicing teachers. Correlations were run between total scores generated by the two instruments, scores were assigned percentile ranks and compared, and descriptive statistics were computed for TADS and FPMS scores.

Sample and Procedure

Observations were conducted in eight schools (four high schools, one middle school, and three elementary schools) in Dade County. All grade levels were represented by at least two teachers, and subject area representation closely paralleled that of the FPMS norming study (Appendix M). About 45 percent of observations occurred in elementary classes. Average lesson length was somewhat greater than in the FPMS norming study, and was trimodal in nature, with elementary lessons generally 30 minutes or less in duration, middle school lessons about 40 minutes, and high school lessons about 55 minutes.

All observers were qualified to observe for the Florida Master Teacher Program. A total of 40 observers (20 Tads, 20 FPMS) took part in the study. A total of 76 observations of practicing teachers were conducted using both TADS and FPMS evaluation instruments. Teachers were observed simultaneously by one rater using TADS, and one observer using FPMS. Most teachers were observed twice, on two different days by four different observers/raters. Logistic factors ultimately reduced the number of teachers having four total observations (two TADS, two FPMS) to 66.

Results and Discussion

Nature of the Distributions

Score distributions for the two instruments in this study approximate distributions found in the Master Teacher Program. FPMS scores approximate the normal, with mean, median and mode falling at about the same point, and minimum skewness (.189).

TADS scores distributed in a more skewed fashion, with substantial difference between mean and median scores, and a moderate negative skew (-.58) (Appendix M).

Relationship of TADS and FPMS Scores

The correlation between TADS and FPMS total scores was r=.49 (n=66, P <= .01). This is a positive and significant relationship.

Rank Ordering

The Master teacher standard was set by the State Board of Education at the 75th percentile. To test the overlap between TADS and FPMS scores at this point, scores of the observations from this study (N = 66) on both instruments were assigned percentile ranks, and compared. Table 47 shows that of the 17 individuals scoring at or above the standard on the TADS, 10 of them (58 percent) would also have reached the same standard on the FPMS. Of the 18 individuals reaching the standard on the FPMS, 10, or 55.5 percent would also have reached standard on the TADS.

TABLE 47

Number of Individuals Falling Above 75th Percentile Cuts

	75TH PERCENTILE CUT SCORES FOR TADS	
	ABOVE 75TH PERCENTILE	BELOW 75TH PERCENTILE
	N	N
75TH PERCENTILE CUT SCORES FOR FPMS		
ABOVE 75TH PERCENTILE	10.00	8.00
BELOW 75TH PERCENTILE	7.00	41.00

Limitations

Some concern should be mentioned regarding the general nature of TADS score distributions. The ceiling effect and extreme negative skew may cause two general problems:

1. reduction in levels of relationship between this instrument and other measures, and
2. reduction in the instrument's ability to discriminate at the upper end of the scaling.

The original design called for the observation of 80 teachers (40 elementary, 40 post-elementary) simultaneously by TADS raters and FPMS observers. In practice, only 66 of the intended teachers were observed twice by TADS observers, and twice by FPMS observers. In addition, on

the third day of observation, observations and ratings were not conducted during the same lesson, meaning that teachers were not observed teaching the same lesson on both pairs of observations.

Summary of Findings

1. There is a significant ($r=.49$, $n=66$, $P < .01$) positive correlation between a teacher's scores on TADS and the FPMS summative observation instrument.
2. Applying the 75th percentile standards, FPMS and TADS included the same candidates in the upper quarter approximately 55 percent of the time.

Results from these studies will guide the development of predictive validity studies designed to answer question three stated above -- How do FPMS and TADS scores relate to various measure of student outcome? Predictive validity studies for this purpose are planned for 1985-1986.

BIBLIOGRAPHY

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972) Robust Estimates of Location Survey and Advances. Princeton: Princeton University Press.

Baldwin, L., Medley, D.M., and MacDougall, M.A. (1983) A comparison of analysis of covariance to within-class regression in the analysis of non-equivalent groups, Journal of Experimental Education.

Blom, G. (1958) Statistical Estimates and Transformed Beta Variables. New York: Wiley and Sons.

Burstein, L., Linn, R.L., and Capell, F.J. (1978) Analyzing multilevel data in the presence of heterogeneous within-class regressions, Journal of Educational Statistics, 3:4, 347-383.

Domains: Knowledge Base of the Florida Performance Measurement System. (1983) Florida Coalition for the Development of a Performance Measurement System, Office of Teacher Education, Certification, and Inservice Staff Development, Tallahassee, FL.

Glass, G.V. (1982) Meta-Analysis: An approach to the synthesis of research results. Journal of Research in Science Teaching. 19:2, 93-112.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92:2, 490-499.

Hedges, L.V., and Olkin, I. (1983) Clustering estimates of effect magnitude from independent studies. Psychological Bulletin, 93:3, 563-573.

Jackson, G.B. (1980) Methods for integrative reviews. Review of Educational Research, 50:3, 438-460.

Light, R.J. and Pillemer, D.B. (1984) Summing Up the Science of Reviewing Research, Harvard University Press: Cambridge, MA.

McDonald, F.J. (April 1976). A Report on the Results of Phase II of the Beginning Teacher Evaluation Study: The Effects of Teaching Performance on Student Learning. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Medley, D. (1985). From correspondence and discussions.

Micceri, T. (July 1985) FPMS Validity Report #1: Overview of Florida Performance Measurement System, Predictive Validity Studies A and B; Theoretical Background, Results, and Implications. University of South Florida, Tampa, FL.

Micceri, T. (July 1985) Validity Report 1-A: FPMS Validity Study A, A Short Term Study of Teacher Interns in Elementary Classes. University of South Florida, Tampa, FL.

Micceri, T. (June 1985) Validity Report 1-B: FPMS Validity Study B, A One Semester Study of Experienced Post-Elementary Teachers in Social Studies/History Classes. University of South Florida, Tampa, FL.

Micceri, T. (1983) A Report on the Reliability Study for the Florida Beginning Teacher Evaluation Summative Instrument. University of South Florida, Tampa, FL.

Micceri, T. (1984) Technical Report on the Normative Study for the Florida Performance Measurement System Summative Instrument. University of South Florida, Tampa, FL.

Ragosa, D., Floden, R.E., and Willett, J.B. (1984) Assessing the Stability of Teacher Behavior: Research Series No. 141. Institute for Research on Teaching, East Lansing, MI.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. Journal of Consulting and Clinical Psychology, 51:1, 4-13.

Rosenthal, R. (1978) Combining results of independent studies. Psychological Bulletin, 85:1, 185-193.

Rosenthal, R. and Rubin, D.B. (1982) A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74:2, 166-169.

Soar, R. (1984). From personal conversations.

Sternberg, R.J. (1984). What should intelligence tests test? Implications of a triarchic theory of intelligence for intelligence testing. Educational Researcher, 13:1, 5-15.

Teacher Evaluation Internship Project. (1984) Final Report. University of South Florida, Tampa, FL.

Teacher Evaluation Project. (1983). A Study of Measurement and Training Components Specified in the Management Training Act, Final Report for 1982-1983. Florida Coalition for the Development of a Performance Measurement System, Office of Teacher Education, Certification, and Inservice Staff Development, Tallahassee, FL.

Teacher Evaluation Study: Norming Study, Generic Competence Revision, Feedback Proposal, Cognitive Examinations, and Specialized Domains. (1984). Report for 1983-1984. Florida Coalition for the Development of a Performance Measurement System, Office of Teacher Education, Certification, and Inservice Staff Development, Tallahassee, FL.

Appendix A

Beginning Teacher Program: Directions for Interviewer

The interview of the beginning teacher should cover four areas: identifying data, the beginning teacher's understanding of the Beginning Teacher Program, the procedures that were followed, and the content of the help the beginning teacher received. The interview format includes some questions to guide the information-gathering through these four areas, but it also provides room for additional comments.

Briefly explain your role as data collector for the evaluation of the Beginning Teacher Program. Inform the beginning teacher of their role as a sample subject in the field test of the interview which means their suggestions for interview questions and/or interview changes, as well as suggestions about the Beginning Teacher Program are welcomed. Emphasize that all compiled data will be anonymous and confidential.

Give them a specific statement like the following to start off the interview: "I wish to emphasize that all information gained from this interview will be anonymous and confidential and only considered a part of grouped data. No individuals will be singled out. You were selected as part of a random sample of beginning teachers for this interview."

Give the teacher an overview of what areas will be covered and indicate your willingness to add information, particularly in the summary, so they know everything will be covered and are less likely to digress.

Where appropriate, circle numbers that represent the beginning teacher's response.

This identifying information should be filled out by the interviewer prior to the interview.

I. Demographic information: Are you a representative beginning teacher in our sample of teachers?

1. Social security number _____
2. Grade level of teaching _____
3. Subject(s) if appropriate _____
4. Name of degree institution
   If missing on role, fill in _____
   If out-of-state, give state _____
5. Number of years teaching experience prior to current year _____
6. Sex _____
7. Ethnic group

a. White, non-hispanic
b. Black, non-hispanic
c. Hispanic
d. American Indian, Polynesian
e. Other

8. Have you been signed off for completing the 24 competencies?
   a. Yes
   b. No

Interview

II. Orientation: What preparation for the Beginning Teacher Program did you receive?

1. Did your district provide information about the legal requirements of the BTP? (180 days, 24 competencies)
   1. Yes
   2. No
      a. If yes, who provided this information?
         1. Building level administrator
         2. Peer teacher
         3. Subject specialist
         4. Other, name, role or title _____
      b. How was this information provided?
         1. Workshop, lecture or meeting
         2. Handout
         3. Informal conversation
         4. Other, name, _____

2. Did your district provide information about the requirements of your local school district for the beginning teacher?
   2.1 With regard to: School board policies?
      1. Yes
      2. No
   2.2 School board approved procedures for the BTP?
      1. Yes
      2. No
   2.3 Timelines for the BTP procedures?
      1. Yes
      2. No
   2.4 The appeal process for teachers who are unsuccessful in completing the BTP?
      1. Yes
      2. No
         a. If yes, who provided this information?
            1. Building level administrator
            2. Peer teacher
            3. Program coordinator
            4. Other, name, title _____
         b. How was this information provided?
            1. Workshop, lecture or meeting

2. Handout
3. Informal conversation
4. Other, how? _____

3. Were you made aware of the Florida minimun essential competencies?
   1. Yes
   2. No
      a. If yes, were the competencies reviewed and explained?
         1. Yes
         2. No
      b. Were they made available in writing?
         1. Yes
         2. No

4. Did the district familiarize you with the terms for elements of the BTP?
   1. Yes
   2. No
      a. If yes, were you made aware of which of the following terms?
         1. Professional Development Plan (PDP)
            a. Yes
            b. No
         2. Florida Performance Measurement System (FPMS)
            a. Yes
            b. No
         3. Portfolio
            a. Yes
            b. No
         4. Summative Instrument
            a. Yes
            b. No
         5. Formative Instrument
            a. Yes
            b. No
         6. Support Staff/Team
            a. Yes
            b. No
         7. Peer Teacher
            a. Yes
            b. No

5. Was your support staff/team identified?
   1. Yes
   2. No
      a. If yes, when were you told the names of your support staff?
         1. First week
         2. Second week
         3. Other: _____
      b. Who served as the other professional educator? (name the role or title)
         1. Classroom teacher
         2. Curriculum coordinator
         3. College educator
         4. Other: _____

III. Implementation: What procedures were followed in your school/district.

1. Were there pre-observation conferences?
   1. Yes
   2. No
      a. If yes, who attended these conferences?
         1. Building level administrator
         2. Peer teacher
         3. Other educator
         4. Other, name role _____
      b. Were competencies or domains to be observed identified during conferences?
         1. Yes
         2. No
      c. Were you consulted about a convenient time?
         1. Yes
         2. No
2. Have you been told when you would be observed?
   1. Yes
   2. No
      a. What observation instruments were used?
         1. Formative 1 2 3 4 5 6 (circle all that apply)
         2. Summative, first screening
         3. Final summative
         4. Other, name _____
      b. Who made the first screening observation (summative instrument)?
         1. Building level administrator
         2. Peer teacher
         3. Other educator
         4. Other, name _____
      c. When was the first screening observation made (summative instrument)?
         1. First 4 weeks
         2. 5 - 6 weeks
         3. 7 - 9 weeks
         4. After 9 weeks
      d. If you started the BTP in August or September, how many times have been observed?
         1. _____
3. Was a post-observation conference held?
   1. Yes
   2. No
      a. If yes, how many observations were conducted before a post-observation conference?
         1. One
         2. Two
         3. After all observations were over
         4. Other, name _____
      b. Who attended the post-observation conference? (circle all that apply)
         1. All support staff members
         2. Building level administrator
         3. Peer teacher

4. Other, name _____

c. How many conferences were held before all competencies were verified (sign-off)?
   1. Only one
   2. Two-three
   3. More than three
   4. Estimated number _____

4. Was a Professional Development Plan written as a team?
   1. Yes
   2. No
      a. If yes, were data from the observations used in developing the plan?
         1. Yes
         2. No
      b. Were you aware of the problems of teaching that the plan was designed to help you solve?
         1. Yes
         2. No
      c. Did you have input in developing the plan?
         1. Yes
         2. No
      d. Was the Professional Development Plan revised or kept up-to-date after each observation?
      e. How specifically were your needs identified? (the most common level)
         1. By domain (i.e., need help in lesson organization, planning, management of conduct, etc.)
         2. By concept (i.e., need help in maintaining academic focus)
         3. By indicator (overlapping, higher order questions)
      f. How soon after the assignment were you able to complete assigned activities?
         1. Within 2 weeks
         2. Within 1 month
         3. More than 1 month
      g. What kind of activities were assigned for instructional improvement and who directed your activity? (enter director number from list below)
         1 = Building level administrator
         2 = Peer teacher
         3 = Supervisor
         4 = College based educator
         5 = Other

Director#		Rank
_____	Workshop - focused	_____
_____	Workshop - general	_____
_____	Consultation/demonstration	_____
_____	Guided observation of skilled teachers	_____
_____	Independent work	_____
_____	Specific instruction by team members	_____
_____	Professional reading	_____

What? _____

_____

_____

_____

h. Rank order your 3 most useful experiences above (put rank in right hand column above).

IV. Interviewer's comments:

1. Most difficult questions for teacher to answer? _____

_____

2. Questions in which information is questionable (list numbers)?

_____

3. General comments: _____

_____

4. Length of interview _____
5. Name of interviewer _____

Appendix B

Domain 10.0 Conferring with Teachers Observation Instrument

OBSERVATION INSTRUMENT

Domain 10.0 Conferring With Teachers

Florida Performance Measurement System
Coalition for the Development of a Performance Evaluation System
Office of Teacher Education
Certification and Inservice Staff Development
Tallahassee, Florida

This instrument is designed to record effective and ineffective observable indicators of Consultant behavior in Domain 10.0 identified by the Florida Coalition.

DIRECTIONS:

1. All of the information on the instrument must be complete to constitute a valid observation.

2. Observe and code the entire session from beginning to end (minimum of a twenty minute observation).

3. Check the time the observation begins and ends and record the length of the observation on the instrument.

School_____ School Number_____

District_____ District Number_____

Observer's Signature_____

Date of Observation_____ Time_____ AM__ PM__

Consultant's Signature_____

INSTRUCTION FORMAT/FRAME FACTORS

Consultant's Name_____
(Print) (last) (first) (middle)

Consultant's SS# __ __ __ - __ __ - __ __ __ __

Circle Consultant's Sex
1. Male 2. Female

Circle Consultant's Racial/Ethnic Group
1. White, Non-Hispanic 2. Black, Non-Hispanic
3. Hispanic 4. Other

Observer's Name_____
(last) (first) (middle)

Observer's SS# __ __ __ - __ __ - __ __ __ __

Length of Observation in minutes_____

1. Circle the Consultant's instruction classification
1. Principal 2. Supervisor
3. Peer Teacher 4. Other (specify)_____

2. Circle the type of conference observed
1. Summative Observation 2. Formative Observation
3. Other (specify)_____

3. Circle level of identified student population
1. Elementary 2. Middle/Junior High
3. Senior High 4. Adult

FLORIDA PERFORMANCE MEASUREMENT SYSTEM

DOMAIN 10.0 CONFERRING WITH TEACHERS

CATEGORY	INDICATORS	FREQUENCY	FREQUENCY	INDICATORS
10.1 RAPPORT	1. Begins conference promptly			1. Delays conference
	2. Provides orientation/review			2.
	3. Displays facial/body behavior congruent with discussion			3. Displays incongruent facial/body behavior
10.2 SKILLS 10.3 PROBLEM FORMULATION	4. Examines observation data with teacher/gives feedback			4. Elicits speculative remarks (rather than using data)
	5. Elicits teacher concern/interpretation of data			5. Expresses opinions, gives information, assumes agreement
	6. Asks specific questions one-at-a-time			6. Asks non-specific/multiple questions
	7.			7. Makes non-specific responses (I see, uh huh)/criticizes
	8. Uses specific reinforcement			8. Uses general reinforcement
	9.			9. Uses the same technique repeatedly (rigidity, mental set)
	10.			10. Solicits personal information/allows irrelevant content
	11. Rephrases teacher comments			11. Does not respond to what teacher says
	12. Summarizes analysis of data/emphasizes			12. Does not summarize analysis
	13. States the problem			13. Switches problem focus abruptly / rushes into planning
	14. Solicits restatement of discussion to confirm problems			14. Does not confirm discussion
10.4 DEVELOPMENT OF PLANS	15. Explores alternatives, eliciting teacher's ideas			15. Lays out plan without collaboration
	16. Induces teacher to summarize plan			16.
	17. Confirms plan			17. Does not confirm plan
	18. Establishes time for follow-up			18.

Appendix C

Relationship of Different Student Strata with Posttests

Tables 48 and 49 show the simple correlations between the absolute number of students in one of the academic strata (high, average and low) and posttest scores for each of these strata. These results indicate that the more lower achieving students in a class, the lower the slope of the regression line between pre and posttests, thus indicating reduced gains for other subgroups (average and higher achievers). The more high achieving students, the greater the slope of the regression equation.

This result however can easily be misleading, as the greater the number of lower achievers in a class, the lower the pretest mean tends to be. To test this, pretest means for lower achievers were entered first into a regression model, followed by the absolute number of lower achievers, using raw posttests as the dependent variable. The results for Study A were not significant, (Rsq ch = .01, F ch = .882, P <= .52), but for Study B they were significant (Rsq ch = .097, F ch = 4.69, P <= .05). Therefore, even in view of the lower pretest scores caused by more lower achievers, the impact of a greater number of lower achievers is negative on all lower achievers.

TABLE 48

Correlations Between EGS and Number of Students in Groups

	Expected Gain Scores			
	All Students	High Achievers	Average Achievers	Low Achievers
Total Number of Students in a Class	0.098	-0.082	-0.066	0.094
	P= 0.49	0.59	0.66	0.51
	n=50	n=44	n=46	n=49
Number of High Achievers in a Class	0.322	0.213	0.171	0.243
	P= 0.03	0.16	0.26	0.11
	n=44	n=44	n=44	n=43
Number of Average Achievers in a Class	-0.060	-0.176	-0.038	-0.076
	P= 0.68	0.25	0.79	0.61
	n=46	n=44	n=46	n=45
Number of Low Achievers in a Class	-0.396	-0.303	-0.303	-0.249
	P= 0.00	0.04	0.04	0.08
	n=49	n=43	n=45	n=44

TABLE 49

Correlations Between EGS and Number of Students in Strata

Study B

	Raw Posttest Scores			
	All Students	High Achievers	Average Achievers	Low Achievers
Total Number of Students in a Class	0.0017 (39) P=0.496	-0.0188 (38) P=0.455	-0.0635 (39) P=0.351	0.0060 (36) P=0.486
Number of High Achievers in a Class	0.4780 (38) P=0.001	0.2993 (38) P=0.034	0.2732 (38) P=0.049	0.3312 (35) P=0.026
Number of Average Achievers in a Class	0.0235 (39) P=0.444	-0.0667 (38) P=0.345	-0.0809 (39) P=0.312	0.1086 (36) P=0.264
Number of Low Achievers in a Class	-0.5628 (36) P=0.000	-0.3105 (35) P=0.035	-0.3878 (36) P=0.010	-0.4417 (36) P=0.003

Appendix D

Identification of Outlier in Study B

The second sample for Study B is reduced by one outlier. The following quote from Andrews et. al (1972) deals with generally accepted techniques used by statisticians to robustify their results and conclusions.

> "A good statistician (or research worker) never uses the mean, even though he may honestly claim to do so. In reality, he will look at the data and set aside outliers and suspected outliers in order to treat them separately; then he will take the mean of the rest. ...And since such a rejection procedure will get rid of distant contamination, any such method will work in that it prevents the worst" (damage to results).

The following two plots indicate the relationship between pre and posttest scores for this teacher, as well as the relationship between the ineffective score and expected gain scores. In both cases, the results are somewhat peculiar. Table 50 shows an extremely unusual ceiling effect, in which all of the lower scoring pretest students attained about maximum on the posttest, yet three of the highest scoring students on the pretest scored considerably lower on the posttest than on the pretest (percentile rank). Table 51 shows the general ceiling effect present for all teachers on the ineffective side of the instrument, and the outlying nature of this one individual's score. In addition, it was suggested that posttests may not have been administered properly in this class. For these reasons, analyses were reconducted without including this individual's scores.

TABLE 50

Plot of Pre and Posttests for Outlier - Study B

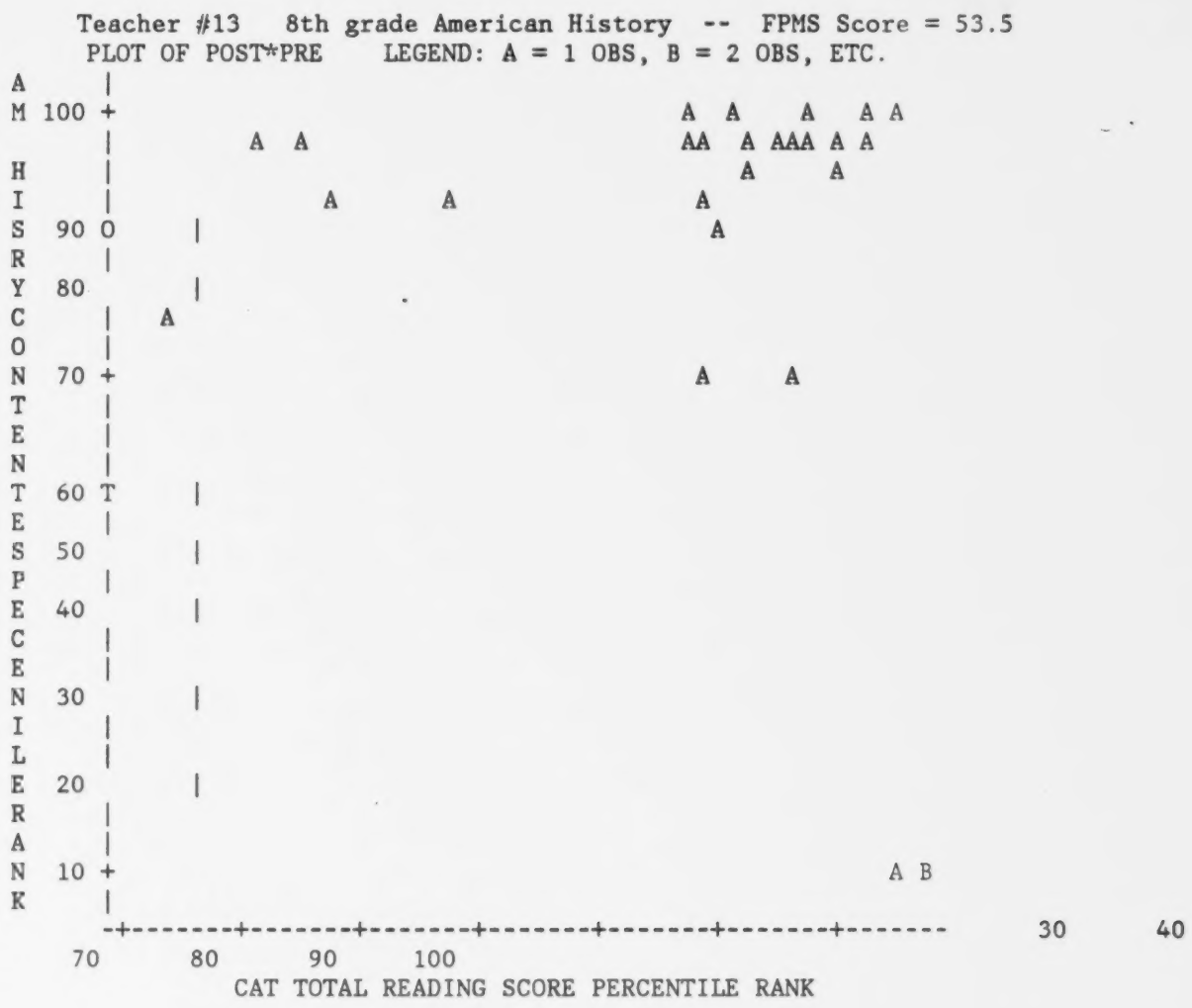
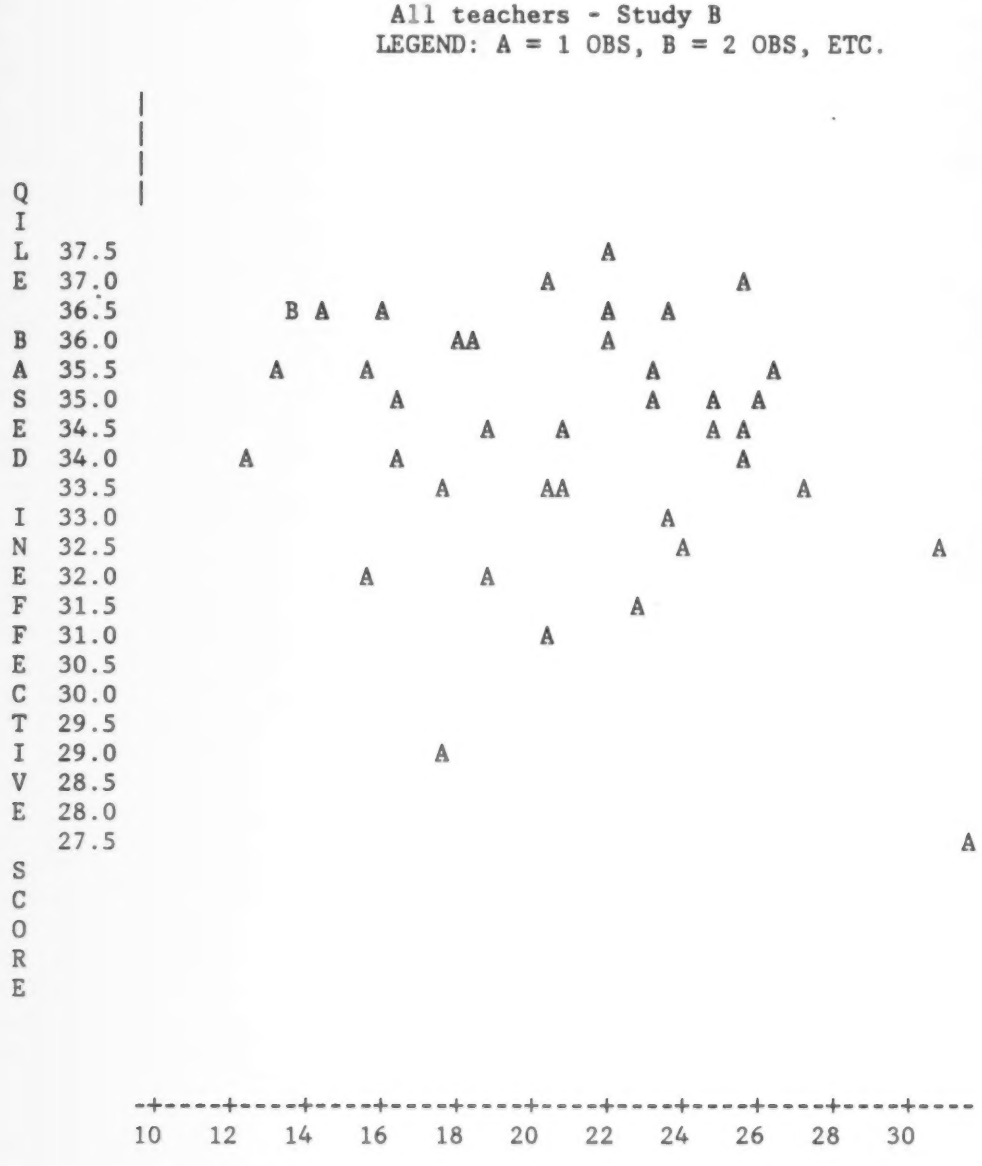


TABLE 51

Plot of FPMS Ineffective Scores and Class Posttest Means



Appendix E

Quartile-Based Scoring Method

In order to develop a scaling for the FPMS summative observation instrument that would be applicable to either individual observations or those conducted in groups, the data base used in the norming study was divided into two randomly selected groups (n=670, and n=553). The following factors were given careful consideration:

1. Item distributions tend to be "J" shaped, almost approximating the exponential. This means they have extremely long, straggled tails. It is generally useful to reduce the effects of such long tails on item scores during the scaling process. For this reason, normal standardized scores were considered as a scaling technique. However, they were not appropriate for this application, as such scores are not generalizable to groups other than the one in which they are used.

2. Although the underlying assumptions of this instrument dictate that more coded behaviors among different indicators result in better instruction, this does not necessarily mean that "more is better" for any specific behavior. In some cases, "more" may well be "worse" (Soar, 1984). For this reason, it is necessary to limit the influence on a total score resulting from any specific item on the instrument. This avoids the problem of overweighting that results when more frequently occurring behaviors are weighted equally against less frequently occurring behaviors by frequency or percentage.

In view of the above considerations, it was decided to use the semi-interquartile range as a basis for scaling items. Thus, for each effective item, individuals who exhibit a frequency of behaviors equal to or above the top 25 percent as related to the normative data base receive a score of three; those who fall between the 25th and 75th receive a score of two; and those who fall into the bottom 25th percent receive a score of one point.

It is not expected that each teacher score in the top or even the middle 25 percent on every item, as each lesson and topic may require a slightly different mix of behaviors for effectiveness. However, the total distribution of item frequencies across all behaviors will indicate the activity of a teacher during a lesson. By summing scores, both positive (effective) and negative (ineffective), it is possible to relate a teacher's behavior to the entire norming sample without any one item overly influencing a teacher's standing.

Appendix F

Scaling and Instrument Development - Design and Results

Research Questions

The scaling questions of interest for this study include:

1. Is the quartile-based scaling equal to or better than the expected normal score scaling (Blom 1958) in predicting student outcome scores?

2. Do different items/indicators predict student outcomes to different degrees?

3. Do some items/indicators operate in a curvilinear fashion in the prediction of student outcomes?

4. Do some subscores predict student outcomes better than others?

5. Is a total score (containing both effective and ineffective scores) appropriate as a predictor of student outcome, or are subscores better predictors? (This question bears directly on the Master Teacher Program as total scores are being used to identify the top 25 percent of teachers in classroom performance.)

6. Would revised scores, based on item analysis, factor analysis, or content analysis predict student outcomes better than the current total and subscores?

7. Do relationships between FPMS scores and student outcome measures differ across grades and frame factors?

8. How can one best identify clusters of behaviors that may work together in effecting either greater student achievement or student task engagement?

All of these questions must be addressed during the predictive/criterion period of the FPMS instrument development (two to five years). Such questions could result in scale or instrument revisions, or even the development of new instruments to supplement the current summative and formative instruments of the FPMS.

Five studies have been, or are being, conducted relating FPMS scores with various student outcomes. These allow the researcher to develop

post hoc scale revisions, test their power on the developmental sample, then cross validate their relative power against future samples from different studies. This is an extremely powerful, yet rarely applied approach to scale development.

The Creation of Subscales and Revisions

Given the preceding considerations, several analyses and scale development procedures were implemented. A number of subscales of the FPMS summative instrument and several methods of scoring the instrument were tested. The following three subscales receive primary consideration in these tests:

1. selected questioning items,

2. selected feedback items, and

3. selected affective items.

The summative instrument contains the following information:

1. The absolute number of effective and ineffective behaviors exhibited during a lesson across all indicators.

2. The frequency of occurrence of each specific indicator both effective and ineffective.

3. The relative frequency of effective and ineffective behaviors either

   a) item by item, or

   b) for a total effective vs ineffective relationship.

Different scalings use, or fail to use, the varying types of information available from the instrument. For example, one consideration when using the total frequency of item occurrence is the possibility of overweighting more frequently occurring items. This problem may be alleviated by limiting the absolute weight assigned to any specific item by its frequency.

Questioning and Feedback Subscales

As Ragosa et. al (1984) suggest, it is possible to develop Bernouli type percentage data from various observation systems. Two areas within the FPMS summative instrument were selected for this purpose:

1. Questioning -- teacher questions as defined by the following four indicator categories:

   a) item 5 effective-lower order questions

   b) item 6 effective-higher order questions

   c) item 24 ineffective-multiple questions/unison response

   d) item 25 ineffective-nonacademic/procedureal questions

2. Feedback -- teacher feedback as defined by the following item/indicator categories:

   a) item 8 effective lower order questions

   b) item 7 effective higher order questions

   c) item 26 ineffective feedback, and

   d) item 27 general praise.

From this may be computed the percentage of effective and ineffective questioning and/or feedback items identified in a lesson. Total scores, consisting of the percentage of effective questioning/feedback minus the percentage of ineffective questioning/feedback were created. In addition, percentages of effective high order questioning, low order questioning, and ineffective questioning and feedback may be computed. The formulae to compute such scores are:

HIGH ORDER PCT = item 6 / (5+6+24+25)

LOW ORDER PCT = item 5 / (5+6+24+25)

INEFFECTIVE PCT= items 24+25 / (5+6+24+25)

TOTAL QUESTIONING = ((5+6)/(5+6+24+25)) - ((24+25)/(5+6+24+25))

To better match these percentage scores with the quartile-based scaling generally applied to the FPMS, the percentage of effective minus ineffective questions was divided into ten categories as shown below.

Percentage Score	Revised 9 Point Score
.01 to .09	0
.10 to .19	1

.20 to .29	2
.30 to .39	3
.40 to .49	4
.50 to .59	5
.60 to .69	6
.70 to .79	7
.80 to .89	8
.90 to .99	9

This revised scaling was found to correlate at least as well as the original Bernouli type scores with student outcome measures.

The total questioning subscore proved to be useful in predicting student outcomes in Study A, and was added to total scores for the instrument for two revisions (one and two), which were applied in subsequent studies.

Affect Variables

Another area of concern and interest is that of teacher "affect", which investigates the nature of the relationship established between a teacher and her class. The affect variable created for this purpose contains three indicators (two effective and one ineffective) of the FPMS summative instrument that appear to tap this construct:

1. Item 17 - Teacher exhibits enthusiasm,

2. Item 18 - Teacher exhibits positive body behaviors

3. Item 37 - Teacher frowns, deadpan, lethargic

Each of these items provides a general indication of the "atmosphere" in a classroom. A teacher exhibiting positive affective behaviors should develop a different classroom climate than one using negative or neutral behaviors. Through the use of these indicators, a relationship between teacher affective behavior and student outcomes may be studied. In order to test the results of this affect variable, these three items were removed from the total score and entered into a regression model with the reduced total score in predicting task engagement. They were then entered into a regression model using achievement posttest scores as the dependent variable, and task engagement, affect, and the reduced FPMS score as independent variables. A second regression model used task engagement as the dependent variable with affect and the reduced FPMS score as independent variables.

Table 52 clearly shows the lack of relationship between the affect variable and either achievement or engagement. This is interesting, for it becomes the only subscore of the FPMS instrument not to relate to task engagement.

TABLE 52

Relationships of Affect Variables to Student Outcomes

	Partial Correlations With Student Outcomes Scores for Average Students (mid 50%)			
	Study A		Study B	
	N	Affect	N	Affect
Achievement	49	.11	40	- .15
Task Engagement	49	- .19	40	.09

Scoring Revisions

Four scoring revisions were tested to determine their relationship with measures of student outcome.

Revision one. In this revision, all four questioning items (5,6,24,25) were deleted from the quartile score, and replaced by the revised questioning subscale (range 0 to 9).

Revision two. This revision included all effective items unless they had a non-zero negative correlation (r = .15) with a student outcome measure, in which case, the ineffective indicator was included. The final form of this revision included the following twenty three items: 1, 4, 6, 7, 9, 10, 11, 13, 14, 16, 17, 18, 20, 21, 22, 23, 27, 30, 31, 33, 35, 38 and 39.

Revision three. In this revision, the percentage of ineffective scores was subtracted from the percentage of effective scores. This was done at the total score level as follows:

$$\text{Revision three} = (E / Tot) - (IE / Tot)$$

where:

Tot = total teacher behaviors
E = effective teacher behaviors
IE = ineffective teacher behaviors

Revision four. This revision is computed by first summing the coded behaviors for a given item. The percentage of ineffective behaviors for the item is then subtracted from the percentage of effective behaviors. Then, to include the absolute number of effective behaviors for an item, but not to over include them, this score is multiplied by the square

root of the absolute number of effective behaviors. Scores for each item are summed to achieve a total score for a lesson. By summing across all possible indicators, total scores ranging from a low of -99 to a high of +99 are possible.

Since some behaviors lack either an effective or an ineffective indicator, all effective only indicators will score as the square root of their occurrence, and all ineffective only items will score as the negative of the square root of their occurrence.

Results of Scale and Subscale Findings

All revisions except revision four, developed in Study A, produced higher correlation coefficients with student outcomes than the present quartile scaling method although none of these differences even approached significance (see Table 24, page - in text). The fact that both revision one and three showed higher coefficients with achievement measures than the quartile measure supports the increased attention given questioning in the revisions.

For all scalings, task engagement correlates higher with FPMS than does student achievement, with math classes showing a consistently higher correlation than social studies classes. Only the ENS scaling for social studies produces a correlation coefficient lower than .27; and if one were to compute the median coefficients for all scalings from Table 24 in the body of the report they would be: Study A - .55 for task engagement (.50, .54, .56, .58, .68), .38 for math, and .325 for social studies, all of which, if considered separately, are significant at the .10 level. For Study B median correlations are .60 for task engagement and - .02 for achievement.

Revision three, being an extremely simple approach to scaling, and apparently a powerful one, may prove most appropriate for general evaluation purposes. However, it is probably most inappropriate for Master Teacher purposes because of the potential reticence of observers to code ineffective behaviors in evaluation contexts.

Feedback and affect variables show little if any relation to either student engagement or achievement.

Relationship Among FPMS Scalings

The relationship between each scaling is important to determine whether the scalings appear to be measuring the same thing. The preceding table, identifying the information contained in each possible scaling, suggests that Revision One should correlate highly with the Quartile scaling, that Revision Three will probably not correlate highly with any of the other scalings, and that the ENS scaling should correlate best with the quartile-based scaling.

Table 53 displays the interscale correlation for each pair in Study A. It appears that the predictions in the preceding paragraph are fulfilled by the actual correlations.

TABLE 53

Correlations Among Various FPMS Scalings for Study A

	ENS n=51	QUAR-TILE n=51	REVISION ONE n=51	REVISION TWO n=51	REVISION THREE n=51	REVISION FOUR n=51
ENS SCORE	1.000 P=0.00					
QUARTILE SCORE	0.810 P=0.00	1.000 0.00				
REVISION ONE	0.733 P=0.00	0.935 0.00	1.000 0.00			
REVISION TWO	0.719 P=0.00	0.906 0.00	0.978 0.00	1.000 0.00		
REVISION THREE	0.625 P=0.00	0.738 0.00	0.834 0.00	0.809 0.00	1.000 0.00	
REVISION FOUR	0.789 P=0.00	0.930 0.00	0.939 0.00	0.918 0.00	0.824 0.00	1.000 0.00

APPENDIX G

STUDENT SERVICES: OBSERVATION INSTRUMENT

Florida Performance Measurement System
Coalition for the Development of a Performance Evaluation System
Office of Teacher Education, Certification and Inservice Staff Development
Tallahassee, Florida

State of Florida
Department of Education
Tallahassee, Florida
Ralph D. Turlington, Commissioner
Affirmative action/equal opportunity employer

The observation instrument is designed to record effective and ineffective observable indicators of behavior in student services domains identified by the Florida Coalition.

DIRECTIONS:

1 All of the instruction format and frame factor information on the instrument must be complete to constitute a valid observation.

2 Observe and code the entire session from beginning to end (minimum of twenty minute observation). The applicant must have the responsibility of leading the session.

3 Record school name and number; district name and number; and starting time and ending time (if the session does not occur in a school setting, record N/A in the school name and number spaces)

4 Check the time the observation begins and ends and record the length of the observation under the frame factor.

School _____ School Number _____

District _____ District Number _____

Observer's Signature _____

Date of Observation _____ Starting Time _____ ☐AM ☐PM Ending Time _____ ☐AM ☐PM

Applicant's Signature _____

MT-18

INSTRUCTION FORMAT/FRAME FACTORS

Applicant's Name (PRINT) _____ (last) (first) (middle)

Applicant's SS# ___ ___ ___ ___ ___ ___ ___ ___ ___ DOE # _____

Circle applicant's sex
1. Male 2. Female

Circle applicant's racial/ethnic group
1. White, Non-Hispanic 2. Black, Non-Hispanic
3. Hispanic 4. Other

Observer's Name (PRINT) _____ (last) (first) (middle)

Observer's SS# ___ ___ ___ ___ ___ ___ ___ ___ ___

Length of observation in minutes _____

1. Circle the applicant's instructional classification.
1. Counselor 2. Occupational Specialist
3. Psychologist 4. Social Worker
5. Other (specify) _____

2 Circle the type of session observed.
1. Counseling 2. Consultation
3 Other (specify) _____

3 Circle the nature of the participants.
1 Students 2. Faculty
3 Parents 4. Other _____

4 Circle level of identified student(s)
1. Elementary 2. Middle/Junior High
3. Senior High 4. Adult
5. Multi level

5-6 Complete the information required in the following items

5. _____ Total number of participants

6. _____ Applicant's number of complete years student services experience

STUDENTS SERVICES: Observation Instrument						
			FREQUENCY		FREQUENCY	
MANAGEMENT	1	Begins promptly		1.		Delays session
	2	Uses materials/space in an orderly manner		2.		Loses time due to use of materials/space
	3	Conducts orientation--purpose/limits/duration		3.		Conducts incomplete orientation
	4.	Maintains momentum/problem solving focus		4.		Loses momentum, over-dwells, interruptions, talk unrelated
	5	Conducts a beginning review--recalls specific events/data		5.		
QUESTION-RESPONSES	6.	Asks questions - single/clarifying/elicits concerns/feelings		6.		Asks multiple questions; forces responses
	7	Uses participant(s)' responses; gives feedback		7.		Ignores participant(s)' response/ignore feelings;expresses sarcasm, disgust/harshness; makes value judgments
	8	Uses specific reinforcement		8.		Uses general reinforcement
USE OF DATA	9	Presents information/identifies or utilizes data		9.		Insufficient use of data
	10	Summarizes data analysis		10.		Moves to problem(s)/concern(s) without data analysis
PROBLEM-PRACTICE	11	Identifies problem(s)/concern(s)		11.		Ignores participant(s)' concerns/inadequate identification of problem
	12.	Provides for practice--role play/simulation		12.		Extends or changes topic(s) without practice
	13	Summarizes problem(s)/concern(s)		13.		Conducts inadequate summary of problem(s)/concern(s)
CLOSURE	14.	Formulates goal(s)/agreements/explores alternatives		14.		States unclear goals/alternatives
	15.	Summarizes/confirms plans		15.		States plans without seeking agreement
	16.	Assigns responsibilities/future expectations		16		States vague responsibilities/expectations
COMMUNICATION	17			17.		Uses loud grating, high pitched, monotone, inaudible talk
	18.			18.		Uses scrambled discourse/vague terms
	19.	Uses congruent body language		19.		Uses incongruent body language
	20	Expresses enthusiasm/interest verbally		20.		
	21	Emphasizes important points		21		

Appendix G

Final Report on the Student Services Instrument-Pilot Study

Executive Summary

Much like the FPMS summative observation instrument, the effective side of this instrument yields a reasonable range of scores, while the ineffective side produces an extreme negative skew, with most individuals exhibiting fewer than three ineffective behaviors. We suggest the use of a total score produced by combining both effective and ineffective scores for purposes of evaluation, as the highest reliability estimates result from this longer instrument.

Reliability estimates are not high, however, this is not surprising given the nature of the study (page 127 and 128 for details). The internal consistency estimate closely approximates that of the FPMS summative instrument (r = .78). Intercoder agreement is probably considerably higher than the estimate (r = .56) shows. Limited variablilty among the twenty eight subjects observed to derive this estimate combined with the multisession nature of the observations both contributed to the low estimate. It is my opinion that a study to be conducted on the observations by several trained observers of the same video tapes will result in considerably better estimates of reliability, and that in reality, the intercoder agreement is fairly high.

The pilot test included 100 observations conducted in a variety of settings. Only one small subset of scores appears unusual. Psychologists consulting in a group session with other faculty members had considerably lower scores than any other subgroup. This small subeffect was strong enough to make group consultation scores appear lower than individual consultation scores, and group scores appear lower than individual scores in general despite the fact that group scores for other subgroups closely approximate individual type session scores.

Reliability Estimates for Observation Instruments

Various Approaches

Numerous approaches may be taken for estimating the various forms of reliability characterizing observation instruments and rating scales. The three most important are:

1. DISCRIMINANT RELIABILIY - the consistency with which a score on an instrument differentiates among a group of subjects,

2. INTERCODER/RATER AGGREMENT - the consistency with which different coders/raters record the same behaviors in a similar fashion for different subjects, and

3. STABILITY - the consistency of a subject's behaviors over time or situations.

Several approaches to obtaining these estimates are popular at the present time, including

1. the generalizability (intraclass) approach,

2. the correlational approach, and

3. numerous inter-rater agreement estimates.

One of the strongest of these approaches is that of generalizability. This form of analysis was applied to the data obtained in this study, however, it must be noted that the resulting coefficients are probably extremely low due to the nature of the data. With different raters/coders observing different subjects at different times, it is impossible to separate the variation due to raters, sessions, subjects, rater X session interaction, rater X subject interaction, and subject X session interaction from general variance. Thus all of this variance must be lumped together in the error term (Shrout, 1978).

Due to the problems noted above, three separate estimates of reliability were conducted on each of the three possible scales for this instrument; total effective score, total ineffective score, and combined total for effective and ineffective scores:

1. Generalizability - twenty eight pairs of observations were run in a oneway ANOVA in which observer X subject, observer X session, and session X subject interaction were all included with error,

2. Coefficient Alpha (a measure of internal consistency - a weak substitute for other measures of Discriminant Reliabiliy), and

3. A correational estimate of intercoder/rater agreement. in which the twenty eight scores of the first observers were correlated with the twenty eight scores of the second observers.

Scaling

After viewing the descriptive data obtained from 100 observations for the Student Services instrument it was decided to use a dichotomous scaling for this instrument. The infrequent nature of behaviors on the Student Services instrument disallows any use of a higher level scaling. There are several possibilities for other scalings within the instrument for future research purposes, however, for Master Teacher purposes, it is inappropriate to apply these. This type of low-level scaling generally reduces the variation and therefore, the ability of an instrument to differentiate between different levels of behavior for different subjects.

Analysis

Apart from the reliability estimates already described, descriptive statistics were generated for all Frame Factor and Context Variables available, as well as for the three possible scales derived from the instrument.

Oneway ANOVAs were run across the following variables identified by the instrument:

1. Session type - consultation or counseling,

2. Grouping type - individual or group,

3. Job classification - counselor, occupational therapist, psychologist or social worker, and

4. Nature of participants - students, faculty, parents, other, and faculty with parents.

Factorial ANOVAs (unweighted means analysis) were conducted to test for possible interaction effects between grouping type and other context variables.

Correlations were run between the three scales of the instrument and the number of participants in group sessions, as well as subjects' experience.

Results

Sample

A sample of 100 observations was obtained from this pilot study, including 28 pairs of observations conducted at two different times by two different observers on the same subject.

The sample consisted primarily of white (81%) females (67%) holding a masters or higher degree (90%). Most observations took place in an urban setting (97%), with each level of schooling represented by at least 20% of the observations (elementary - 24%, middle - 23%, secondary only - 6%, multilevel 46%). The median length of observation was 20 minutes (mean 25 minutes). Observation length ranged from 15 to 60 minutes, with 43 (43%) taking only 15 minutes.

Forty nine observations were made of counselors, four of occupational therapists, twenty one of psychologists, and twelve of social workers. Session types included forty five counseling, twenty nine consulting, and eleven defined as some other type. Thirty eight of the sessions were with indivdiuals while forty were in a group situation. The number of participants in these sessions ranged from one to sixty four, with a median group size of 4.7 (the mean of 11.9 was pulled far out by a few extremely large groups). Thirty eight of these sessions occurred with students, sixteen with faculty members, six with parents alone, four with parents and faculty, and seven with various combinations of individuals.

Descriptive Statistics

Much as with the FPMS summative instrument, considerable variation occurred for the effective side of the instrument, with little or no variation for the ineffective behaviors. Effective total scores ranged from two to 19, with a mean of 12.4 (median 13.1), ineffective total scores ranged from 10 to 18 with a mean of 16.5 (median 17.0 -- note, one score on the ineffective side was four, however, this score was so low, that it was eliminated from analysis, as it would tend to considerably distort both scale estimators and point estimators of location) Combined scores ranged from 10 to 37, with a mean of 28.9 (median 29.4).

BEST COPY AVAILABLE

Reliability Estimates

Table 36 on page 118 in the report shows, the highest reliability estimates result for the total score (combining) both effective and ineffective sides of the instrument. The ineffective side exhibited many zeros, thereby resulting in considerably reduced variance, and an extremely low reliability estimate.

Total score estimates should be used in any evaluation conducted using this instrument.

Frame Factors and Context Variable Effects

Several effects were obvious from preliminary analysis of these data:

1. Counseling is conducted primarily with students (94%), with the remaining two sessions occuring with parents.

2. Consultation never occurs with students, however, may occur with any one or combination of the following groups:

   a) faculty members,

   b) parents,

   c) administrators,

   d) other staff, or

   e) any combination of the above.

Highly significant differences ($p < .001$) exist for the grouping type in which any form of consultation or counseling occur (Table 58). When grouping is crossed with other variables such as type of session, type of counselor/consultant, etc., a significant interaction effect occurs between the grouping and the other factors. Further analysis of this effect shows that low scores tend to arise in only one small subset of cases:

1. that in which psychologists consult with other faculty in a group situation.

In fact, this particular situation accounts for some 50% (six of 13) of the effective scores below 8.0 for the entire sample of 100. At the present time, John Hansen is attempting to determine whether some unanticipated effect such as the psychologist not leading the group occurred to cause this result. Tables 54 and 55 below specify the number of cases and the mean scores for the total effective scores (ineffective scores did not vary across any groups or combination of groups) broken across all context variables identified by this instrument.

TABLE 54

COUNSELING SCORES ACROSS ALL SERVICES KEY VARIABLES

TYPE OF SESSION=COUNSELING

				INSTRUCTIONAL CLASSIFICATION					
		TOTAL		COUNSELOR		PSYCHOLOGIST		SOC WORKER	
		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE	
		N	MEAN	N	MEAN	N	MEAN	N	MEAN
	NATURE OF PARTICIPANTS								
TOTAL	STUDENTS	34	13.44	28	13.54	3	12.33	3	13.67
	PARENTS	2	11.50	.	.	.	.	2	11.50
TYPE OF GROUP	NATURE OF PARTICIPANTS								
INDIVIDUAL	STUDENTS	19	13.47	16	13.19	2	14.50	1	16.00
	PARENTS	2	11.50	.	.	.	.	2	11.50
GROUP	STUDENTS	15	13.40	12	14.00	1	8.00	2	12.50

There is some indication that the grade level of a school may have some slight impact on the effective side of the instrument (Table 56), however, this may be no more than the effect of the size of a school.

The number of non-whites and rural settings was far too small for any analysis in this pilot. These must wait for the analysis of the Master Teacher data itself.

No important or significant relationships resulted from the correlation of any of the three scales with

TABLE 55

CONSULTATION SCORES ACROSS ALL SERVICES KEY VARIABLES

TYPE OF SESSION=CONSULTATION

		TOTAL		INSTRUCTIONAL CLASSIFICATION							
				COUNSELOR		OCC SPECIALIST		PSYCHOLOGIST		SOC WORKER	
		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE		TOTAL EFFECTIVE SCORE	
		N	MEAN	N	MEAN	N	MEAN	N	MEAN	N	MEAN
	NATURE OF PARTICIPANTS										
TOTAL	FACULTY	13	11.23	4	11.25	.	.	5	10.40	4	12.25
	PARENTS	4	12.75	1	15.00	1	10.00	2	13.00	.	.
	OTHER	7	9.14	1	18.00	.	.	5	6.80	1	12.00
	FAC & PARENT	1	16.00	1	16.00	.	.	.	.	.	.
TYPE OF GROUP	NATURE OF PARTICIPANTS										
INDIVIDUAL	FACULTY	7	13.14	1	14.00	.	.	3	13.67	3	12.33
	PARENTS	4	12.75	1	15.00	1	10.00	2	13.00	.	.
	OTHER	2	16.50	1	18.00	.	.	1	15.00	.	.
GROUP	FACULTY	6	9.00	3	10.33	.	.	2	5.50	1	12.00
	OTHER	5	6.20	.	.	.	.	4	4.75	1	12.00
	FAC & PARENT	1	16.00	1	16.00	.	.	.	.	.	.

1. the number of participants in a session (slight negative correlation), or

2. the experience of a subject.

The attached Tables show the univariate relationships among each of the three scales derived from the instrument and the several context variables discussed.

TABLE 56

SCORES ON STUDENT SERVICES INSTRUMENT ACROSS GRADE LEVEL

	GRADE LEVEL OF SCHOOL					
	ELEMENTARY		MIDDLE		ALL LEVELS	
	N	MEAN	N	MEAN	N	MEAN
TOTAL INEFFECTIVE SCORE	14	16.43	9	16.22	24	16.83
TOTAL EFFECTIVE SCORE	14	12.64	9	14.89	24	12.17
COMPLETE TOTAL SCORE	14	29.07	9	31.11	24	29.00

TABLE 57

SCORES ON STUDENT SERVICES INSTRUMENT ACROSS SESSION TYPE

	TYPE OF SESSION					
	COUNSELING		CONSULTATION		OTHER	
	N	MEAN	N	MEAN	N	MEAN
TOTAL INEFFECTIVE SCORE	27	16.52	16	16.75	4	16.50
TOTAL EFFECTIVE SCORE	27	13.85	16	11.38	4	11.75
COMPLETE TOTAL SCORE	27	30.37	16	28.13	4	28.25

TABLE 58

SCORES ON STUDENT SERVICES INSTRUMENT ACROSS GROUP TYPE

	TYPE OF GROUP			
	INDIVIDUAL		GROUP	
	N	MEAN	N	MEAN
TOTAL INEFFECTIVE SCORE	23	16.78	24	16.42
TOTAL EFFECTIVE SCORE	23	14.04	24	11.67
COMPLETE TOTAL SCORE	23	30.83	24	28.08

TABLE 59

SCORES ON STUDENT SERVICES INSTRUMENT ACROSS COUNSELOR TYPE

	INSTRUCTIONAL CLASSIFICATION					
	COUNSELOR		PSYCHOLOGIST		OTHER	
	N	MEAN	N	MEAN	N	MEAN
TOTAL INEFFECTIVE SCORE	32	16.47	10	16.90	5	16.80
TOTAL EFFECTIVE SCORE	32	13.91	10	9.70	5	12.20
COMPLETE TOTAL SCORE	32	30.38	10	26.60	5	29.00

TABLE 60

SCORES ON STUDENT SERVICES INSTRUMENT ACROSS SUBJECT TYPE

	NATURE OF PARTICIPANTS					
	STUDENTS		FACULTY		OTHER	
	N	MEAN	N	MEAN	N	MEAN
TOTAL INEFFECTIVE SCORE	29	16.48	7	16.86	11	16.73
TOTAL EFFECTIVE SCORE	29	13.72	7	12.71	11	10.55
COMPLETE TOTAL SCORE	29	30.21	7	29.57	11	27.27

BIBLIOGRAPHY

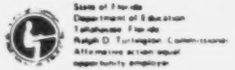
Frick, T., and Semmel, M.I. (1978) Observer agreement and reliabilities of classroom observational measures. Review of Educational Research. 48:1 157-184.

Mitchell, S.K. (1979) Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin 86:2, 376-390.

Shrout, P.E., and Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86:2, 420-428.

LIBRARY MEDIA: DATA COLLECTION INSTRUMENT

Florida Performance Measurement System
Coalition for the Development of a Performance Evaluation System
Office of Teacher Education, Certification and Inservice Staff Development
Tallahassee, Florida



This instrument is designed to record the presence or absence of documented information for the library media domains identified by the Florida Coalition.

DIRECTIONS:

All of the information on both sides of the instrument must be complete to constitute a valid observation.

School _____ School Number _____

District _____ District Number _____

Observer's Signature _____

Date of Observation _____

Applicant's Signature _____

Applicant's Name _____
(PRINT) (last) (first) (middle)

Applicant's SS# ___ ___ ___ ___ ___ ___ ___ ___ ___

Observer's Name _____
(PRINT) (last) (first) (middle)

Observer's SS# ___ ___ ___ ___ ___ ___ ___ ___ ___

19

FRAME FACTORS

1. Circle the type of school.

   1. Elementary
   2. Middle/Junior High School
   3. High School
   4. Other (specify) _____

2. Circle the applicant's primary responsibility.

   1. Print Services
   2. AV Services
   3. ITV Services
   4. Combination of 1 and 2
   5. Combination of 1 and 3
   6. Combination of 2 and 3
   7. Combination of all
   8. Other (specify) _____

3. Circle the race of the applicant.

   1. White Non-Hispanic
   2. Black Non-Hispanic
   3. Hispanic
   4. Other _____

4. Circle the sex of the applicant.

   1. Male
   2. Female

Complete the information required in the following items.

5. _____ Number of students enrolled in school.

6. _____ Number of professional personnel in the library media center.

7. _____ Number of support staff in library media center.

6. _____ Applicant's number of complete years library media experience.

LIBRARY MEDIA DATA COLLECTION INSTRUMENT

FUNCTION	CONCEPT		INDICATORS	EVIDENCED YES	EVIDENCED NO	NUMBER OF
MANAGEMENT	PLANNING	1	Establishes cooperation with school community			YES _____
		2	Establishes goals, objectives, priorities and strategies			
		3	Establishes policies and procedures			NO _____
		4	Develops and maintains budget			
	ORGANIZATION	5	Provides access			YES _____
		6	Arranges Facilities			
		7	Maintains user-oriented environment			NO _____
		8	Arranges and circulates resources			
		9	Manages personnel			
	EVALUATION	10	Conducts program assessment			YES _____
		11	Analyzes data and modifies program			NO _____
COLLECTION DEVELOPMENT	SELECTION/ ACQUISITION	12	Uses approved selection policies			YES _____
		13	Uses aceptable selection procedures			NO _____
		14	Solicits educator and student input			
	MAINTENANCE	15	Provides for maintenance and repair			YES _____
		16	Weeds and inventories collection			NO _____
SERVICES	SKILLS INSTRUCTION	17	Utilizes scope and sequence guide			YES _____
		18	Correlates with classroom instruction			NO _____
	CONSULTATION/ INFORMATION	19	Provides reference assistance			YES _____
		20	Provides motivation and guidance in reading/viewing/listening			
		21	Orients students and educators			NO _____
		22	Instructs and updates educators			
		23	Participates in curriculum development and planning			
	PRODUCTION	24	Provides production resources and technical assistance			YES _____ NO _____
	PUBLIC RELATIONS	25	Promotes program			YES _____ NO _____

Appendix H

Final Report on the Library Media Instrument-Pilot Study

Executive Summary

This instrument appears to fulfill several primary requirements of a rating scale:

1. Acceptable discrimination between individuals,

2. Acceptable interrater agreement

3. Subscale homogeneity, and

4. Unbounded scores (that is, no cases receiving either minimum or maximum possible points on the instrument).

Given the limited nature of the sample included in this pilot, this instrument performs quite well. Scores range widely within identified subgroups (grade level, responsibilities, etc.), while showing very similar mean scores between subgroups.

Background to Analysis

Various Approaches to Reliability Estimation

Numerous approaches may be taken for estimating the various forms of reliability characterizing observation instruments and rating scales. Three of the most important are:

1. DISCRIMINANT RELIABILIY - the consistency with which a score on an instrument differentiates among a group of subjects,

2. INTERCODER/RATER AGREEMENT - the consistency with which different coders/raters record the same behaviors in a similar fashion for the same subjects, and

3. STABILITY - the consistency of a subject's behaviors over time or situations.

Several methods for obtaining these estimates are popular at the present time, including

1. the generalizability (intraclass) approach,

2. the correlational approach, and

3. numerous inter-rater agreement estimates.

One of the strongest of these approaches is that of generalizability. This form of analysis was applied to the data obtained in this study, however, it must be noted that the resulting estimates are probably extremely low due to the nature of the data. With different raters/coders observing different subjects at different times, it is impossible to separate the variation due to raters, sessions, subjects, rater X session interaction, rater X subject interaction, and subject X session interaction from general variance. Thus all of this error must be lumped together in the error term (Shrout, 1978), thereby reducing reliability estimates.

Analyses Conducted

Three separate estimates of reliability were conducted on the data:

1. Generalizability (in this case a very conservative estimate with both variance resulting from between subject effects and variance from rater X subject effects included in the error).

2. Coefficient Alpha (a measure of internal consistency - a weak substitute for other measures of Discriminant Reliabiliy), and

3. A correlational estimate of intercoder/rater agreement.

Scaling

A dichotomous scaling (1=yes, 0=no) is the only appropriate scaling for a checklist type of inventory. The maximum possible score is 25. Despite the fact that this type of low-level scaling generally reduces the ability of an instrument to differentiate between different levels of behavior for different subjects, this instrument obtained excellent internal consistency (coefficient alpha) estimates of reliabilty, and exhibited a considerable range of scores. This bodes well for its use in the Master Teacher Program.

Average scores on this instrument were compared across several frame factors characteristic of these disciplines.

Results

Sample

A sample of 25 pairs of ratings for various library/media specialists was obtained. The scores reported result from the average of a pair of raters for a given subject. The sample was 92 percent female, 88 percent white, and 67 percent masters degree holders. The mean experience was 11.7 years, and ranged from two to 25 years. Ninety two percent of the ratings took place in urban settings. Specialists were spread over grade levels with 10 at elementary, eight at middle school and seven at secondary level.

The staffing of these centers varied, although each had at least one staff member in addition to the one observed, and 57 percent had one or two additional staff members. The largest staff consisted of six individuals. The number of professional personnel ranged from zero to three, with 57 percent having one or less. Most settings had at least one non-professional staff member.

Fully 85 percent of these specialists claim multiple responsibilities, including audiovisual, television, and print services.

Reliability

Table 42 on page 129 of the report indicates the reliability estimates for this instrument are relatively high for the total instrument (.77 to .89), high for the Management subsection (.71 to .81), and low to moderate for the other two possible subscales. In both the Collection/Development and Services section of this instrument the consistency of scores was lower than in the Management section. As with all instruments studied it is best to use a single total score on this instrument for purposes of evaluation.

Items

Interestingly, two items, mends resources (88 percent) and arranges facilities (92 percent) were recorded as yes by almost every rater. Other items ranged from 35 percent to 65 percent. This suggests that subjective evaluation of Library Media centers and personnel have probably been based upon the "look" of effectiveness.

Obtained Average Scores

The total instrument exhibited an excellent range of scores particularly given the small sample size. The total score provides a reasonable approximation to the Gaussian (normal) distribution, with mean and median being very close, and a skew of less than .10. The following table indicates several descriptive statistics for a total score and three possible subscores that may be used for this instrument.

TABLE 61

Descriptive Statistics for Total And Subscale Scores

Score	# Items	n	Mean	Median	Std Dev	Low	High
Total	25	25	13.5	14.4	5.4	5	23
Management	11	25	5.6	5.5	2.5	2	10
Collection Development	5	25	3.2	3.4	1.3	0	5
Services	9	25	4.7	4.9	2.4	1	9

Frame Factors and Item Scores

Although no significant differences across scores occurred for any of the Frame or Context variables analyzed in this study (at alpha .05), there were some slight indications of differences, particularly with regard to the number of professional personnel and the grade level of the specialist. In both cases, the more, the higher the score. Of the many analyses conducted, the only one to approach significance was that of the number of professional staff for the collection/development subscale score ($F = 2.82$, $df=3,21$, $p<.07$). Even these mean scores were curvilinear in nature, as shown in Table 62.

There also were no important or significant correlations between any total score or subscale scores and experience (highest $r = .14$, $n = 25$).

TABLE 62

Number of Nonprofessional Staff Across Subscores

	Number of Nonprofessional Staff			
	0	1	2	3
Mean Col/Develop Subscore	2.21	3.77	3.50	2.75

BIBLIOGRAPHY

Frick, T., and Semmel, M.I. (1978) Observer agreement and reliabilities of classroom observational measures. Review of Educational Research. 48:1 157-184.

Mitchell, S.K. (1979) Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin 86:2, 376-390.

Shrout, P.E., and Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86:2, 420-428.

Appendix I

Content Validation Process and Results

Content Validation Procedures

The first Addendum to the Coding Manual With ESE Examples was development by the Special Education Task Force, chaired by Mrs. Laverne Graves, Bureau of Education for Exceptional Students. The Task Force members represented several ESE categorical areas and developed examples for the purpose of describing typical ESE teacher behavior according to the FPMS Indicators. Meeting four times during the summer and fall of 1984, this task force generated the first addendum draft, pooling their knowledge on the research base and its applications in all types of exceptional student education.

The document is intended to be generic, for all categories of exceptionality, rather than specific to only one or two types of classrooms. The task force tried to retain as much of the language and intent of the original FPMS as possible, while providing examples of effective and ineffective indicators that were appropriate for ESE teachers.

Next, the Addendum was then reviewed by Donovan Peterson, Jean Borg, Betty Fry, Catherine Morsink, and Diane Fardig. This review considered appropriate placement of the examples in keeping with the purpose and form of the original coding manual. Based on this review the Addendum was revised and sent to 225 ESE practitioners for content validation review. The practitioners were recommended by their exceptional education directors and represented all ESE categorical areas in small, medium, and large Florida school districts. Practitioners were requested to rate the 100 Indicator examples on a 5-point scale (1=low; 5=high) that reflected the following:

1. CLARITY means the degree to which the example is a clearly logical explanation of the effective or ineffective behavioral indicator.

2. ADEQUACY means the degree to which the example contains descriptors consistent with the effective or ineffective behavioral indicator.

3. APPROPRIATENESS means the degree to which the example is representative and reflective of the effective or ineffective indicator.

Practitioners were asked to slash out any item that was completely unclear, inadequate, or inappropriate.

One hundred twenty practitioners responded to the request for content validation of the Addendum. The responses were analyzed by University of South Florida personnel. The results stabilized at about 20 respondents. These results formed the basis for revision of the Addendum examples. Any example having at least two descriptors scoring below 4.0 were considered for revision.

Twenty-four respondents wrote letters and/or comments expressing concern or reaction to the example. These comments were considered in the revision process for the items having two or three descriptors below a 4.0 rating.

In no case was there a revision of the coding manual format. Revisions to the Indicator examples included editorial changes, modifications when an item was unclear, additional descriptors when an item was inadequate, and in a few cases deleting an item that was inappropriate. In several cases examples and notes from the original coding manual were replaced. The purpose and intent of the FPMS, and related Domains references were important considerations in the revision process.

Of the 100 Indicator examples reviewed, 79 items had at least two descriptors scoring above 4.0 and no item below 3.50. Twenty-one examples had at least two descriptors scoring below 4.0. The following section provides a description of the 21 examples, a summary of ratings and comments, and the change (if any) made in the example.

Indicator 1. Delays

1.B.1 --- allows student to engage in non-directed or unsupervised activities.

1. Clarity:	3.94	3. Appropriateness:	3.91
2. Adequacy:	4.06	4. Found in original:	No

Summary of comments:

Several reviewers slashed non-directed; apparently teachers view non-directed activities as positive, part of students becoming independent.

Change:

In order to clarify that this is an ineffective example, the statement was edited as --- allows students to engage in unsupervised activities or activities that have no apparent direction.

1.B.3 --- does not routinize and handle with dispatch activities such as roll reports and lunch tickets.

1. Clarity:	3.88	3. Appropriateness:	3.88
2. Adequacy:	3.85	4. Found in original:	Yes

Summary of comments:

Several reviewers indicated that certain ESE teachers are not required or expected to perform these duties, including therapists, resource

teachers, and itinerant teachers.

Change:
No change in example. A note was added as follows:
Note: some ESE personnel (e.g., therapists, itinerant or resource teachers) may not be expected to handle such activities as roll reports and lunch tickets.

1.B.4 --- chates with students so that classwork is delayed.

1.	Clarity:	3.94	3.	Appropriateness:	3.61
2.	Adequacy:	3.71	4.	Found in original:	Yes

Summary of comments:
Reviewers indicated strongly that this activity may be necessary to clarify issues from the previous day, deal with a problem that occurred in another class, or set the tone for the regular class. Several stated that this communication was essential. Teachers indicated that omitting this activity might delay classwork.

Change:
In order to clarify that this is an ineffective example, the statement was edited as: --- engages in inappropriate or purposeless chatting with student(s) so that classwork is delayed.

Indicator 2. Handles materials in an orderly manner.

2.A.2 --- manages auxiliary personnel in setting ;up and in the monitoring of the safe use of special equipment, during instruction.

1.	Clarity:	3.79	3.	Appropriateness:	3.71
2.	Adequacy:	3.68	4.	Found in original:	No

Summary of comments:
Reviewers indicated that the example reflected what would be seen in certain classrooms such as those for the physically impaired, or severely handicapped, but not in classrooms for the mildly handicapped. Teachers also noted that not all ESE classrooms have auxiliary personnel.

Changes:
a. An example from the original coding manual for this effective indicator was replaced as follows:
--- maintains business-like procedures for housekeeping activities such as passing out papers and folders, use of audiovisual equipment and supplementary materials.
This example more adequately represents activities found in classrooms for the mildly handicapped.
b. The example was clarified in this manner:
--- manages auxilliary personnel (e.g., teacher aides, volunteers, student interns) in setting up and in the monitoring of the safe use of special equipment, during instruction.
c. The notes from the original coding manual were replaced.
Note 1. Normal use of chalkboard is not tallied.
Note 2. Tally each use of material following the intervening-rule episode. See page 2 of this manual.

Indicator 2. Does not organize or handle materials systematically

2.B.1 --- does not implement special equipment/devices

1. Clarity: 4.03
2. Adequacy: 3.49
3. Appropriateness: 3.17
4. Found in original: No

Summary of comments:

As with example 2.A.2., reviewers indicated that the example reflected what would be seen in certain classrooms such as those for the physically impaired, severely/profoundly mentally handicapped, or visually impaired. The teachers stated that this example would not be appropriate for the mildly handicapped who, as a group, do not require special adaptive equipment.

Changes:

a. An example from the original coding manual was replaced (to represent teachers of the mildly handicapped) as follows:
   -- interrupts classwork in order to get materials, such as portable chalkboard, projector, books.
b. The example was clarified as:
   -- does not implement special equipment/devices when needed.
c. Replace Note 3 as follows:
   -- Loss of instructional momentum due to poor use of materials is tallied here.

2.b.2 --- does not manage auxiliary personnel in effective use of special equipment.

1. Clarity: 3.79
2. Adequacy: 3.41
3. Appropriateness: 3.32
4. Found in original: No

Summary of comments:

As with examples 2.A.2 and 2.B.1, reviewers indicate this description does not represent all ESE teachers.

Change:

Add a NOTE as follows:

Note . In some cases teachers will have no auxilliary personnel (e.g., no teacher aide).

Indicator 3. Orients students to classwork/maintains focus on assigned task.

3.A.3 --- Provides advanced organizers verbally and visually.

1. Clarity: 3.50
2. Adequacy: 3.72
3. Appropriateness: 3.84
4. Found in original: Yes, partially, as "... provides advance organizers"

Summary of comments:

Apparently reviewers, who are practitioners and not FPMS trained, may not have understood the term "advanced organizers."

Change:

Since the principals using the Addendum will be FPMS trained, no change was made in this example.

3.B.1 --- accepts questions or requests from the student that are unrelated to assigned task.

1. Clarity: 4.27
2. Adequacy: 3.88
3. Appropriateness: 3.55
4. Found in original: Rephrased

from "accepts nonacademic questions or requests from students."

Summary of comments:
Reviewers' comments were similar to those made regarding example 1.B.1; this activity is perceived as necessary, related to good communication, and helpful in maintaining interest in the subject.

Change:
The example was intended to describe talk/activities that are unrelated to the subject of instruction, therefore the example was edited as:
--- accepts questions or requests from the student that are inappropriate, purposeless, or unrelated to the assigned task.

Indicator 5a. Asks single factual questions
Editorial change: Replaces NOTES 1, 2, and 3 as in the original coding manual, page 7.
5.b .Asks questions that require analysis/reasons
5.c.1 --- asks students to demonstrate ability to combine two previously learned skills to form a new one.

1.	Clarity:	3.66	3.	Appropriateness:	4.00
2.	Adequacy:	3.83	4.	Found in original:	No

Summary of comments:
Reviewers comments were varied. Some teachers felt that the example was at too high a level for certain students including trainable and severely/profoundly mentally handicapped. Other reviewers noted that this example is "hard to observe." Some stated that the example did not require analysis; a check of the dictionary definition of analysis confirms this.

Change:
Since there are two other ESE examples of this Indicator as well as a note, the example was deleted.

Indicator 5.b Poses non-academic questions/non-academic procedural questions
5.d.1 --- asks questions that call for personal opinion or that are answered from personal experience.

1.	Clarity:	3.73	3.	Appropriateness:	3.03
2.	Adequacy:	3.43	4.	Found in original:	Yes

Summary of comments:
Reviewers strongly indicated that this example is effective in the special education setting. Some reviewers felt that personal opinion would be used to relate to "real-world applications" (5.C.3).

Change:
--- asks questions that call for personal opinion or that are answered from personal experience and are unrelated to the task at hand.
5.2.2 --- asks nonacademic procedural questions such as "How do you fold your papers?" "Does everyone know how to exchange his/her library book?"

1. Clarity: 3.42
2. Adequacy: 3.29
3. Appropriateness: 3.03
4. Found in original: Yes

Summary of comments:
As in 5.D.1., reviewers indicated that this example is effective in the special education setting.

Change:
--- asks nonacademic procedural questions that can be categorized as outside the content to be covered.

5.D.3 --- asks questions that call for general non-academic information such as "How many of you got all the questions right?"

1. Clarity: 3.81
2. Adequacy: 3.38
3. Appropriateness: 3.25
4. Found in original: Yes

Summary of comments:
See comment summary for 5.2.2

Change:
--- asks questions that call for general nonacademic information and that do not move the classwork along.

Indicator 6. Ignores student or response/expresses sarcasm, disgust, harshness.

6.B.2 --- speaks angrily.

1. Clarity: 4.06
2. Adequacy: 3.97
3. Appropriateness: 3.89
4. Found in original: Yes

Summary of comments:
Most reviewers who wrote comments generally approved the example.

Change: None.

Indicator 8. Provides for practice.

All examples were validated here. A definition and examples of prescriptive practice were added for the users of the Addendum as follows:

Note 1. Prescriptive practice is practice that matches students' learning needs. For example, students with limited short-term memory may need spaced rather than massed practice; or practice with concrete manipulative objects. Other students may be visual learners and need practice that uses a visual approach. Likewise, auditory learners need practice that includes an auditory component.

The following item, inadvertently placed in Indicator 9 (ineffective) was replaced in Indicator 8 (ineffective) where it was originally intended:

--- provides uncontrolled practice or practice without instruction.

Indicator 10. Remains at desk/circulates inadequately

10.B.1 --- assigns a task to a student but does not monitor activity

or tells student to see him/her if assistance is needed.

1. Clarity 3.91
2. Adequacy: 3.79
3. Appropriateness: 3.68
4. Found in original: Rephrased from ". . . makes an assignment and leaves the students to work it out themselves, or tells them to see her/him if they need assistance."

Summary of comments:
Several reviewers stated that if the students' goal is to ask for help when needed, this would not be an ineffective indicator. Teachers provided examples of tasks that would challenge students in this manner.

Change:
To increase appropriateness, a note was added corresponding to the item as follows:
Note 1. If teacher's goals for developing student independence include completing tasks without being monitored and to ask for assistance when needed, this would not be tallied.

Indicator 11. Gives definitions or examples only
11.B.2 --- gives one or more examples or nonexamples of a concept but fails to state either a definition or attributes; accepts the same treatment of a concept by student.

1. Clarity: 3.94
2. Adequacy: 4.06
3. Appropriateness: 3.97
4. Found in original: Yes

Summary of comments:
Some reviewers questioned " . . . accepts the same treatment of a concept by students." This phrase may have been confusing.

Change:
--- gives one or more examples or non-examples of a concept but fails to state either a definition or attributes. Accepts incomplete treatment of a concept by students.

Indicator 14. Develops criteria and evidence for value judgement
14.A.2 --- gives or engages the students in giving a rating or judgment of an object (thing, event, person, or whatever) as the end product of the evaluative procedure (example: This apron won first place in the home economics fair. Why? It seems to be very sturdy. See how easily I can put it on. And it covers me so I'll keep clean. The sewing was very good.).

1. Clarity: 4.00
2. Adequacy: 3.94
3. Appropriateness: 3.97
4. Found in original: Rephrased from ". . . gives or engages students in giving a rating or judgment of an object (thing, event, person, or whatever) as

the end product of the evaluation procedure.

Summary of comments:
Most of the reviewers who made comments had no stated problem with this example.

Change:
None in example. Note 1 was added by DOE personnel for users:
Note 1. Helping low functioning students develop values may take the form of liking or disliking, rather than the actual formation of value judgments.

Indicator 17. Uses vague/scrambled discourse
17.2 --- states ideas and then restates them over and over in an effort to be clear.

1. Clarity:	3.94	3. Appropriateness:	3.76
2. Adequacy:	3.76	4. Found in original:	Yes

Summary of comments:
Some reviewers indicated that they saw no difference between this item and "emphasizes important points" or "draws attention to key points by repeating them throughout lesson." Others indicated that this behavior may be necessary with some ESE students (e.g., those with memory deficits). The other examples in Indicator 17 were validated.

Change:
Example deleted.
17.4 --- uses words that are indiscriminate or do not specify extent or degree, e.g., "some," "little," "much," and "the rest."

1. Clarity:	4.00	3. Appropriateness:	3.71
2. Adequacy:	3.83	4. Found in original:	Yes

Summary of comments:
Most of those reviewers who wrote comments had no stated problem with this example. It is possible that some teachers use these words with lower functioning students (e.g., trainable mentally handicapped).

Change:
--- uses words that are indiscriminate and vague (e.g., "some," "little," "much," "the rest") when words that specify extent or degree are appropriate to the functioning level of students.

Indicator 20. Stops misconduct
20.A.3 --- can demonstrate effectiveness of intervention with appropriate documentation.

1. Clarity:	3.76	3. Appropriateness:	4.03
2. Adequacy:	3.94	4. Found in original:	No

Note of explanation:

This item was incorrectly placed here; it belongs in the planning domain.

Change:
Deleted example.

Most of the changes described here were based on the results of the content validiation of the Addendum to the coding manual. Other changes were made in order to maintain the original coding manual format and directions (e.g., replacing certain examples and notes). In no instance was an Indicator statement changed.

Appendix J

Rating Checklist for Awarding Additional Points

Instructions

Think of receiving this checklist and being responsible for identifying and verifying additional points for the Master Teacher Program. Then,

1. Rewrite the directions if they are not clear.

2. Delete indicators you consider of little or no value by marking through them.

3. Correct the wording of unclear or ambiguous indicators.

4. Add indicators you think are of value, but not represented on the checklist. All indicators must be instruction related and generic for all instructional personnel.

5. Add categories you think are of value, but are not represented.

6. Rate each category by circling a number from 1 to 6 -- not important at all to very important -- in determining superior performance of a teacher.

7. Try to simplify the checklist, not make it more complicated.

Note:

1. Do not add categories or indicators that are on the summative observation instrument.

2. The rating scale of 1 to 6, not at all important/very important, will not appear on the checklist used for awarding additional points. It only appears on this form for purposes of selecting items administrators consider important in rating superior performance.

Data from Administrators on Rating Categories for Additional Points.

Check the appropriate items

Type of School:

_____ High School
_____ Junior High School
_____ Middle School
_____ Elementary School

Number of Teachers:

_____ 0 to 25
_____ 26 to 50
_____ 51 to 100
_____ 101 to 200
_____ 201 plus

School District _____

School District Number _____
Teacher-Candidate's Name _____
Teacher-Candidate's Social Security Number _____

Master Teacher Program
Checklist for Awarding
Additional Points for Teacher Performance

Directions:
Below is a list of categories and their corresponding indicators. Check each observed indicator you consider as evidence of superior teacher performance of that category. If you have no evidence of performance of any indicator, do not check.

	Very Important					Not at all Important
Category 1. Knowledge of Subject Matter	6	5	4	3	2	1
a) Teacher shows evidence of superior knowledge of instructional service/activity for which s/he is responsible.						
b) Teacher presents the content accurately, or corrects errors if made.						
c) Teacher demonstrates knowledge of developments in the subject field and does not present out-dated information.						
d) Teacher presents subject matter suitable to the grade level he/she teaches.						
e) _____						
Category 2. Knowledge of Oral and Written Language	6	5	4	3	2	1
a) Teacher uses correct syntax in oral and written discourse.						
b) Teacher demonstrates correct spelling or publicly corrects student's misspelled words.						
c) Teacher demonstrates command of language skills by expressing ideas effectively, using correct pronounciation and vocabulary suitable to the instructional level.						
d) _____						
Category 3. Knowledge of Computational Skills	6	5	4	3	2	1
a) Teacher uses the correct skills.						
b) Teacher does not make errors in computations, or						

corrects them when made.

c) _____
_____
_____

Category 4. Fosters Intellectual Skills 6 5 4 3 2 1

a) Teacher performance that provides experience and practice in thinking with data in problems that involves direct and inverse proportion.
b) Teacher behavior that provides experience and practice in the use of samples in problems where there are no pure cases.
c) Teacher behavior that provides experience and practice in solving problems that involve relations among two or more variables.
d) Teacher behavior that provides experience and practice in solving problems that require the use of controlled conditions for a solution.
e) Teacher performance that gives experience and practice in the examination of arguments; examination that requires, among other things, identification of premises, logical validity tests, and analysis of major terms.
f) _____
_____
_____

Category 5. Intercultural Education 6 5 4 3 2 1

a) Teacher presents content that communicates respect for cultural differences.
b) Teacher makes use of resources that show contributions from various cultures.
c) Teacher accepts contributions from students, parents, and others and objectively relates them to appropriate instruction.
d) _____
_____
_____

Category 6. Evaluation (items from Domain 6.0) 6 5 4 3 2 1

a) Teacher tells students the purposes of the test, specifies test content, and tells students how to take tests.
b) Teacher sets up physical arrangements (seating, ventilation, lighting) favorable to optimum student performance on examinations.
c) Verbal behavior of the teacher that sets a warm and encouraging atmosphere for the examination.
d) Teacher uses test data to identify and correct errors and at the same time encourages students in their work.
e) Teacher uses test data to identify where students need special help.

f) _____
_____
_____

Category 7. Workroom Atmosphere 6 5 4 3 2 1

a) Teacher allows or encourages student self-direction and independent exploration.
b) Teacher gives students opportunity to select activities related to their interests.
c) Teacher encourages optimum student involvement in class activities throughout class period.
d) Teacher verbal expression that is free of slang and crude expressions.
e) Teacher exhibits appropriate social behavior with respect to social standards in the presence of students.
f) _____
_____
_____

Category 8. Interpersonal Relationships with Students 6 5 4 3 2 1

a) Teacher recognizes all students without regard to their varied backgrounds; is fair, consistent, courteous and respectful and conveys impartiality to the student, regardless of handicaps, sex, ethnic origins, or ability level.
b) Teacher behavior that induces minorities and culturally different students to participate in class activities.
c) Teacher gives corrective assistance in a constructive manner showing awareness of ethnic differences, socioeconomic backgrounds, handicaps, etc.
d) Teacher helps the student who is ridiculed or demeaned by other students.
e) _____
_____
_____

Category 9. Teacher Expectations 6 5 4 3 2 1

a) Teacher holds each student accountable for the quality of work that will increase his or her academic self-concept.
b) Teacher holds each student accountable for the highest quality of work appropriate for his/her ability.
c) _____
_____
_____

Category 10. Physical Environment 6 5 4 3 2 1

a) Teacher keeps a neat classroom or laboratory and

appropriately adjusts lighting and temperature within control of the teacher to optimum performance of activities.

b) Teacher arranges workroom furniture to facilitate work and maintains neat and orderly storage of materials.

c) Teacher uses relevant workroom equipment -- media, up-to-date bulletin boards, displays, learning centers, etc., -- to support and/or enrich learning experiences.

d) Teacher selects and recommends proper equipment, and cares for and maintains it.

e) Teacher makes students aware of the proper and safe use and care of instructional facilities and equipment.

f) _____
_____
_____

Category 11. Planning for Instruction (Domain 1 items) 6 5 4 3 2 1

a) Teacher's presentation or lesson plan shows evidence of selection of content appropriate to the curriculum, learner's needs, and grade level.

b) Teacher selects appropriate materials to provide clarity.

c) The execution of activities that demonstrate the teacher's skill in sequencing, pacing, and organizing instruction.

d) _____
_____
_____

Category 12. Individualization 6 5 4 3 2 1

a) Teacher activities that result in the formation of groups for different purposes, e.g., to adjust instruction to rates of learning, modes of learning, and learning difficulties.

b) Teacher behavior that results in the selection and use of a variety of activities to meet individual needs.

c) Teacher maintains records of his or her students' performance on diagnostic and other tests and uses data therefrom to plan and conduct instruction.

d) _____
_____
_____

Category 13. Uses Multiple Media 6 5 4 3 2 1

a) Teacher activities that entail the use of computers, films, videotapes, and audiotapes in instructional activities.

b) Teacher activities that use diagrams, maps, and photographs, etc. to clarify and emphasize main points in a lesson.

c) Teacher behavior that emphasizes and clarifies relationships among variables by the use of graphs

and tables.

d) _____
_____
_____

Category 14. Innovativeness and Independence 6 5 4 3 2 1

a) Teacher behavior that indicates willingness to take action on his or her judgement.
b) Teacher behavior that indicates ability to respond positively to suggestions for improvement.
c) Teacher behavior that shows maturity, emotional stability, energy, good health and poise in ordinary work-day situations.
d) Teacher performance that exhibits varied teaching procedures and techniques.
e) _____
_____
_____

Category 15. Professional Conduct 6 5 4 3 2 1

a) Teacher works harmoniously with team teachers, peers, and support staff.
b) Teacher specifies duties and gives necessary instructions to support paraprofessional activities with neither negative effects nor interference with workroom program.
c) Teacher does not engage in criticism of peers, administrators, or parents, especially in the presence of students.
d) _____
_____
_____

Category 16. Professional Growth 6 5 4 3 2 1

a) Teacher voluntarily attends meetings of professional pedagogical and subject area groups.
b) Teacher presents evidence of current voluntary work in inservice activities, university courses, and other training opportunities.
c) Teacher demonstrates new skills and knowledge from reading and other self-improvement activities.
d) _____
_____
_____

Category 17. Counsels with Students 6 5 4 3 2 1

a) Teacher behavior that establishes a working relationship with an individual student.
b) Teacher behavior which enables the participants to specify and analyze the problem.

c) Teacher supplies necessary information or facilitates the development of a plan to resolve the identified problem.
d) Teacher provides opportunity to help student evaluate his/her progress or the resolution of the problem.
e) Teacher behavior that shows respect for a student's right to privacy.
f) _____
_____
_____

Category 18. Peer Collaboration 6 5 4 3 2 1

a) Teacher freely shares data, information, and opinions in conference with peers.
b) Teacher participates in a cooperative effort to plan implementation of the agreed-upon resolution of a problem.
c) Teacher freely accepts his or her share of responsibility for carrying out plans developed cooperately with peers.
d) _____
_____
_____

Category 19. Management 6 5 4 3 2 1

a) Teacher schedules facilities to meet the specific needs of instructional activities, e.g., library, media, practice rooms, etc.
b) Teacher reserves time for use of special equipment.
c) _____
_____
_____

Appendix K

Review of the Florida Performance Measurement System

Dr. Patricia M. Kay, Baruch College of New York
Dr. Carolyn Evertson, Peabody College of Vanderbilt
Dr. Harold Mitzel, Pennsylvania State University

May, 1984

State Board of Education Rule 6A-4.45(4)(c) requires that a three member team review each performance measurement system to be used in the Florida Meritorious Instructional Personnel Program and to evaluate the system for approval or disapproval based on specified criteria. In accordance with the requirement, a team consisting of Carolyn M. Evertson, Harold E. Mitzel, and Patricia M. Kay have reviewed the Florida Performance Measurement System (FPMS) as submitted by the Florida Coalition for the Development of a Performance Measurement System. In reviewing the system, the team studied extensive documentation for the criteria specified in the rule, interviewed the system developers, and conferred on May 21-22, 1984 to conclude deliberations.

The team unanimously recommends the approval of FPMS for use in the Florida Meritorious Instructional Personnel Program. Furthermore, the team makes this recommendation with enthusiasm because of the enormous potential of FPMS for having positive effects on education well beyond its use for measurement in the merit program. The careful use of FPMS holds the promise of significant and dramatic improvement in classroom practice, substantive change in both preservice and inservice teacher education, and refinement and extension of knowledge about effective teaching practice. The extent to which the system has already been accepted and used within the state, coupled with the acclaim with which it has been met by knowledgeable members of the national education community herald the fulfillment of that promise.

The remainder of this report examines several premises upon which FPMS is based, summarizes the team's review for each of the criteria, and concludes with recommendations that the team considers to be most important to the continuing refinement and use of the system.

Premises

Several premises underlying the development of FPMS warranted examination and comment by the team. These premises are important because they point the way toward the limitations of the system.

First, and perhaps most notable, the system is based solely on what is known about demonstrably effective teaching performance: those teaching behaviors that have been shown through research to be associated with or able to bring about desired student learning and/or conduct. In this respect, the system represents the single most comprehensive attempt that the team is aware of to translate the body of knowledge into instructional practice.

Second, instructional behavior is the sole focus of the system. FPMS does not attempt to measure the interaction of the teacher and others (i.e., students or parents) nor does it atttempt to assess personal traits and characteristics of teachers. The system addresses most directly what it is that teachers do to instruct effectively. The system differs from other instructional measurement systems, the use of which make it possible to obtain measures of, for example, the extent to which students are engaged in on-task behavior. In this respect, the system is in accord with established principles of measurement that point to the desirability of direct over indirect measurement.

Third, with the FPMS there is no established or implied sequence or order in which the identified instructional behaviors are to be either performed or observed. The only issue is the extent to which they are present (on the left side of the summative instrument) or absent (on the right). This is consistent with the literature which supports the notion that variability in instructional behaviors is desirable but that is generally silent on specific sequences of activities. Furthermore, this lack of specified sequence results in a system that provides for broad flexibility in individual teaching styles.

Fourth, the sytem addresses competencies at a generic rather than subject or grade specific level. The underlying premise is that there is a single set of teaching behaviors that constitutes effective teaching over all subject areas and grade levels. There is fairly widespread agreement on this issue within the education community and the development and acceptance of the system to date support it. Additionally, it is a notion that appeals to lay people. Because FPMS reflects this idea, its use is likely to enhance communication about what constitutes effective teaching performance not only among educators, but also among interested lay persons.

Criteria Specified in Rule 6A-4.45(4)(c)3.

Valid Knowledge Base

Knowledge base for all elements of the FPMS is an assembled body of information on effective and ineffective instructional performance that is drawn from research on teaching. The knowledge base is comprehensive in scope. The possibility exists that there may be gaps in the knowledge base but these gaps would be owing to a lack of knowledge -- because researchers have not addressed some areas -- and not to a lack of thoroughness in incorporating what is currently known.

The knowledge base is organized into six domains, has been scrutinized by practicing educators, and has been submitted to nationally recognized authorities in the field of teacher effectiveness research for content validation. Their conclusions along with those of the team are that the knowledge base is extremely sound and significantly job-related.

Behavioral Indicators

Each of the behavioral indicators may be traced from its source in the research literature to the summative instrument, thus exemplifying sound measurement practices that lead to content validity. The clarity and specificity with which the behaviors have been defined have been demonstrated in the use of the system. Observers have used FPMS after relatively short training periods and have closely agreed on what was observed. Neither brief training nor close agreement would be possible if the indicators were unclear or inadequately specified.

Normed Performance Observation Instrument

Content Validity.

The FPMS items constitute a comprehensive sample from the domains of the knowledge base. While there is not a strict one-on-one correspondence between the domains of the knowledge base and the domains as represented on the summative instrument, the team is satisfied that all domains are sufficiently represented in the system: four explicitly and two, which cannot feasibly be addressed explicitly, by implication.

Internal Consistency.

Internal consistency reliability estimates (coefficient Alpha) were computed separately for the two scales of the FPMS. The results (.79 and .71) indicate that the items within each scale are related to each other to a satisfactory degree. Often, reliability estimates are attenuated by low frequencies. It is possible that the lower estimate for one of the scales is due to the fact that the behaviors on the scale are less frequently coded than are those on the scale with the higher estimate.

Stability, Interrater Agreement, Discriminant Reliability.

Reliability estimates computed separately for the two scales using the scaling techniques derived from the norming study indicated to a satisfactory degree for preliminary use of the system that:

1. the same observers code the same behaviors in the same way over time (.92 and .85);

2. different observers code the same behaviors in the same way (.98 and .98);

3. different observers on different lessons (occasions) code the same teachers in the same way (.91 and .84).

Again the lower reliability coefficients in two of the pairs may be the result of infrequently coded behaviors on one of the scales.

It is important to note that these reliability estimates were the result of a study in which data were collected and analyzed over only two times (or two occasions or lessons). Data collected over three or more observations would improve confidence in stability. In addition, inspection of the variance analyses from which the reliability estimates were derived indicates that the system reflects more variance due to teacher-observer interaction for one of the scales than might be desired. This points to a potential for observer bias. This effect perhaps can be minimized in the careful selection of observers for any given observation. In any case, the variance due to teachers was much greater than that due to teacher-observer interaction, which fact leads to some confidence in the ability of the system to reliably and objectively differentiate among teachers who perform differently.

Practicality.

The most convincing evidence of the practicality of the system is the extent to which FPMS has already been developed, refined, and put in place with extremely limited resources. The cadre of people already trained and the level of development of materials to support the system are impressive. That the system represents the joint efforts of a coalition of school districts bodes well for its continuing acceptability by the profession.

Score Norms.

Scoring for the FPMS is based upon a normative study in which observations were completed on a sample of teachers broadly representative of Florida public school instructional personnel. As a result of that study, raw scores (frequency counts) are appropriately adjusted for instructional method. Separate norm scales are available for kindergarden through grade five and for grades six through twelve.

Behavioral Specificity.

(See comments previously made under Behavioral Indicators.)

Freedom from Bias.

Analyses of data collected in the norming study indicate that no meaningful relationships exist between sex, ethnicity, experience, or education level and FPMS scale scores. These data along with evidence of reliability and content validity constitute a strong argument that the system is free of biases and other factors unrelated to meritorious teaching. In point of fact, the procedures used to develop the system exemplify job-related, bias-free test development.

Standardized Procedures for Data Collection, Data Analysis, and Evaluation of Performance Measures

Procedures for all aspects of FPMS are well standardized. These include instructions to candidates, observers, and principals; procedures for the conduct of observations; and centralized procedures for data analysis and scoring.

Provisions for the Preparation and Approval of Reliable Observers

The FPMS is supported by an impressive observer training scheme. Both the trainers of observers and observers themselves must meet specified criteria for approval. These criteria include rigorous standards on cognitive examinations of the teacher effectiveness literature as well as on reliability observations. The trainers of observers must also be observed conducting training master trainers. These efforts are supported by well-developed, comprehensive, written and audiovisual materials that effectively target the training and result in high rates of approval among observer trainees.

Recommendations

While the team enthusiastically endorses the use of FPMS for the merit program, there are certain aspects of the system and cautions about its use that are important to bring to the attention of both the system developers and state education policy makers. These involve (1) the need to beware of falling into semantic traps that could conceivably do injury to both FPMS and the merit program; (2) the choice of observers for any given observation; (3) the need to treat FPMS in regard to both policy and practice as a dynamic rather than static system of performance measurement; and (4) the matter of setting standards (cut scores) for its use.

The use of FPMS in the merit program involves projecting images of teaching and education to a large and important audience. It is highly desirable with a system as ambitious and far reaching as this one is to be overly cautious about the words that are used. It is important to provide as many aids to clear and accurate communication and interpretation as possible.

A case in point is the naming of the two scales variously as "effective" and "ineffective" or as "positive" and "negative." These labels inevitably result in such logical absurdities as "ineffective reliability coefficients" to describe perfectly satisfactory reliability coefficients, or "negative reliability coefficients: to describe a column of positive numbers. Indeed, the "ineffective scale" is, in reality, effective. To some extent, this problem can be overcome by naming the scales with neutral words or expressions and the team strongly recommends that this be done. A suggestion is to label the "effective scale" as the X scale and the "ineffective scale" as the Y scale.

Another notable case has its roots in the state rule where one of the criteria is labeled "Interrater Agreement." It more accurately ought to be "interobserver" or, better yet, "intercoder" reliability. Observers in this system do not rate anything. They objectively observe and code specified behaviors. The label should not convey an inaccurate and potentially unacceptable meaning.

In any case, including as many nomographs and other aids to interpretation as possible ought to enhance clear communication and ultimately lead to fewer problems with the acceptability of the system by both educators and lay persons.

The matter of teacher-observer interactions on the Y scale mentioned earlier in this report is of some concern. It is just this quality that led to the discrediting of rating systems. In essence, finding this interaction means that with FPMS, as with all observation systems, the possibility of "cheating" exists if observers are not neutral to begin with. The system developers have noted that observers are less willing to check Y scale indicators than to check X scale indicators. This does not preclude using the system because it is a factor that can be largely controlled by the manner in which the system is implemented.

Obviously, the system must be used with the highest degree of professional responsibility and ethics. To further control reliability and to avoid loss of credibility, it is strongly recommended that a minimum of three observers be used for each observation and, further that all three of the observers be non-interested parties: people to whom the outcome of the observation cannot conceivably make any difference.

FPMS is currently regarded as a system under development and it should continue to be so regarded. Treating it as a static system, one that is finished and needs no further work would seriously undermine both the spirit of its intent and the quality of its functioning. There are at least two areas that will demand continual undating: the knowledge base and the score norms.

To maintain the validity of the system, it must be revised as the knowledge base develops. Researchers will continue to address areas where there are gaps in knowledge about effective teaching. A side benefit of the use of FPMS is that it is likely to spur research on teaching and result in new knowledge. It is imperative that a process of review and updating be incorporated into the system.

Norms can never be regarded as static. The current norms are sufficiently documented to start with, but as the system is used, it will automatically generate new data that ought to be used to revise the present norms. There are now several categories for which score adjustments are made. As data is acquired, the system should look toward replacing that process by constructing separate norms for those categories. In addition, attention should be given to the possibility of setting up somewhat more logical norm categories (i.e., elementary, middle school, high school rather than K-5 and 6-12). This latter is important because people may find it unacceptable to be compared to a reference group of which they do not believe themselves to be a member.

Perhaps nowhere is the need to treat the system as a dynamic one more important than in the matter of setting standards or cut scores. The best way to establish these standards is to design and conduct predictive validity studies and to determine cut scores on the basis of that research. While the system developers plan to do this, it is likely that some other basis for arriving at standards will be used in the meantime. It is important to keep these decisions open for as long as possible while continuing to collect and analyze observation data. This can only serve to increase confidence in the decisions that are made.

A separate but related matter about the setting of standards is important. There is a good bit of evidence to support the idea that the two scales of the FPMS are not equivalent scales and that cut scores should be set separately for each of the scales. The scales are not conceptually equal. One contains indicators of effective teaching and the other of ineffective teaching, and the items on either side of the scale are not necessarily the opposite of one another. Furthermore, it is not likely that the scores on the two scales will discriminate

equally well between teachers who are effective and those who are not. The X scale scores seem to be better at doing this than the Y scale scores. In addition, it is likely that with training the Y scale behaviors will drop out of teaching repertoires leaving, in effect, only one scale on which decisions will be based. It will be important, therefore, to monitor the standards to continue to insure that they are reasonable and valid. It will be easier to make adjustments and seems to make good sense to set cut scores for the two scales separately.

The recommendations that the team makes in this sections of the report should in no way detract from the high regard in which the FPMS should be held. The recommendations are intended to insure that the system, which is a milestone in education, continues to receive the maintenance and support it deserves. If it is forthcoming, then undoubtedly the strong promise of FPMS for having a significant and dramatic impact on education can be fulfilled.

Appendix L

Review of the Teacher Assessment and Development System

Dr. Patricia M. Kay, Baruch College, University of New York
Dr. Carolyn M. Evertson, Peabody College of Vanderbilt
Dr. Harold E. Mitzel, Pennsylvania State University

September, 1984

State Board of Education Rule 6A-4.45(5)(e)2. requires that a three member team review each performance measurement system to be used in the Merit Teacher Program and evaluate the system for approval or disapproval based on specified criteria. In accordance with the requirement, a team consisting of Carolyn M. Evertson, Harold E. Mitzel, and Patricia M. Kay have reviewed the Teacher Assessment and Development System (TADS-MTP) as submitted by the Dade County Public Schools. In reviewing the system, the team studied extensive documentation, interviewed the system developers, and conferred on September 10-11, 1984 to conclude deliberations.

While the system has much to recommend it (most notably, its acceptibility to the community of teachers and administrators in Dade County), the team has found the TADS-MTP documentation addressing several of the criteria specified in the rule to be either weak or flawed, particularly with regard to implications for long term use. Accordingly, the team unanimously recommends that TADS-MTP be approved for use in the Merit Teacher Program for a period of only one year during which time additional developmental work addressing several of the criteria specified in the rule be undertaken.

The remainder of this report examines the background of TADS development upon which TADS-MTP is based, summarizes the team's review for each of the criteria specified in the rule, and concludes with recommendations that the team considers to be essential for the continuing development and further use of the system for the Merit Teacher Program.

Background of TADS Development

TADS-MTP represents the latest revision (specifically for use in the Merit Teacher Program) of a teacher observation and evaluation procedure

that has been under development in the Dade County Public Schools for several years. An examination of the evaluation of TADS sheds light on the main premise underlying the development of the TADS-MTP. This history leads directly to potential strengths and limitations of the system, and is at the root of the team's recommendations for continued use of it.

During the late 1970's, a joint committee representing professional educators of the Dade County Public Schools and the United Teachers of Dade was appointed by the superintendent to address concerns that both teachers and administrators had with the existing standard teacher evaluation procedure. Over a period of several years and guided by the work of the joint committee, the standard observation/evaluation form and procedures were revised and improved upon through several phases.

The procedures in existence by 1982, which had been negotiated between DCPS and UTD, were reflected in the language of the DCPS/UTD bargaining agreement. In essence, those procedures consisted of a set of categories for judging teachers as either acceptable or unacceptable. No observer training schemes existed, and there were problems with specification of performance indicators, assessment criteria, and interpretation of both the content of the form and rules for making judgments about performance. A series of modifications ensued for the purpose of training observers and orienting teachers to the system. Each subsequent modification of the system has been subject to joint DCPS/UTD approval.

There are two aspects of TADS history that are of particular importance to its use in the Merit Teacher Program. First, TADS started as a rating system for making rather loose judgments within a set of descriptive teaching categories. These judgments were subsequently tightened by specifying decision rules and by developing procedures by which observers could be trained to shade their judgments according to the context in which the observation was being made. This kind of a system differs considerably from one which simply records instances of specified teaching behavior. In a rating system, each kind of teaching behavior is judged as either acceptable or unacceptable and those judgments are summed to arrive at a final score. In a recording system, each kind of teaching behavior is counted as being present or absent and the counts are summed to arrived at a final score. In general, recording systems are regarded as more objective than rating systems because there is less room for error due to idiosyncratic judgment or to "halo" effects.

Second, TADS was originally based solely on the conventional wisdom of the profession: it included teaching categories and behaviors that the community of teachers and administrators in Dade County believed to be best professional practice and was without explicit reference to research on effective teaching. As the system developed, attempts were made to link the categories, behaviors, and decision rules to the research literature (as well as to the list of Florida Generic Teaching Competencies). As a result, TADS-MTP can be considered to be based on both conventional wisdom and research on effective teaching.

The argument supporting this dual base has some validity as a starting point for a system to be used in the Merit Teacher Program. As reflected in the state rule, the most desirable base for a performance measurement system is the body of research about demonstrably effective teaching. However, that body of research is not complete in that it does not address all teaching behaviors in all contexts. In order to build a system which can be used comprehensively to measure the performance of all teachers across all contexts, systems developers must either infer that what is known from the research on some contexts will also be true for other contexts, or they must look to other means to fill in the gaps. In the case of TADS-MTP, the gap was filled by seeking to achieve consensus on the conventional wisdom of the profession.

Looking to the experience and beliefs of the teachers and administrators of Dade County to fill in the gaps left by an incomplete research base is not, in and of itself, a limitation. Rather, this route to system development, involving many members of the profession over a period of several years has inevitably led to an acceptance of the system that is not only widespread, but has also been institutionalized by virtue of its inclusion in the collective bargaining agreement.

The potential limitation involves the continuing validity of such a system. To ensure, maintain, and improve upon the validity of a performance measurement system, it is incumbent upon the developers or sponsors to devise a process whereby newly acquired research evidence about effective teaching can be incorporated into the system; particularly, but not only, evidence concerning those teaching behaviors which have not previously been demonstrated to have an effect on student performance.

The evolutionary and negotiated process inherent in TADS development may be a two edged sword in a continuing search for validity. On the one hand, that process could seriously limit development if the beliefs of Dade County educators have been so strengthened or so institutionalized as to make it difficult or impossible to amend the system (and the beliefs) if and when research evidence to the contrary is presented. On the other hand, if a process can be devised for routinely amending the system on the basis of the continual accrual of research results, then the nature of the TADS developmental process bodes well for its continuing widespread acceptance by the profession.

Criteria Specified in Rule 6A-4.45(4)(c)3.

Valid Knowledge Base

As discussed in the previous section, the knowledge base for all elements of the TADS-MTP is an assembled body of information on effective and ineffective instructional performance that is drawn both from the conventional wisdom of the profession and from research on teaching. A survey of the research literature supporting the components of the system has been conducted. The literature review does not make reference to some of the recent work which might have implications for changes in the system, (e.g., Stallings et al, 1979; Rosenshine, 1983; and Evertson et al, 1982) and should be updated to be considered comprehensive. The TADS-MTP program was submitted to three external consultants who have judged that it reasonably reflects the outcomes of research on teaching as well as current conceptions of effective teaching practice. The review team found the testimonial letters by well-known researchers to be superficial and lacking in the specificity one would normally expect from qualified experts. The letters clearly failed to indicate that the experts engaged in a careful and independent analysis of the TADS instrumentation and the data pertinent to it.

The results of a survey of educators practicing in Dade County supports the notion that the system contains teaching behaviors that are significantly job-related and that can reasonably be expected of teachers who are to be recognized for meritorious performance.

Behavioral Indicators

To those who are not oriented to or trained in the use of the system, the behavioral indicators are not specific enough to result in a clear concept of effective and ineffective teaching as contained in the TADS-MTP knowledge base. Each of the behavioral indicators in TADS-MTP requires a summary inference on the part of the observer. Observers must remember what occurs during an observation, and after the observation period is over, must consider eighty-two items for which "acceptable" or "unacceptable" judgments must be made. The close agreement of observers in this system seems to depend more heavily on the intensive training process for the system's use than on the specificity and validity of behavioral indicators. Both the reliability data and the survey of educators in Dade County confirm that the behavioral indicators of TADS-MTP are clearly stated within the context of the system's use.

Normed Performance Observation Instrument

Content Validity.

(See comments previously made under Valid Knowledge Base.)

Internal Consistency.

An internal consistency reliability estimate computed for TADS-MTP indicates that items are related to each other to a satisfactory degree. It should be noted that the existence of an acquiesence response set (whereby raters display the tendency to respond favorably to all items) could artificially inflate estimates of internal consistency. This kind of a response set is a possibility with TADS-MTP since the format of the instrument makes it easier or more socially acceptable or more giving of the benefit of the doubt for observers to check items as "acceptable." This kind of response set is controlled in part by including some items with negative content. The developers appeared to be unaware of the effects of response set on scale development.

Stability.

The extent to which the same observers code the same behaviors in the same way over time is not directly addressed in the TADS-MTP documentation. Results from a field test (generalizability study) indicate that to some extent, stability can be inferred because of high interrater agreement. From such data, however, it is not possible to distinguish between the variability due to differences among observers (or raters) and the variability due to differences among observations or occasions. The literature indicates that with most, if not all, observation systems, and especially with those requiring summary inferences, there is a general drift away from stability over time. A study should be conducted to assess directly the extent to which this phenomenon occurs. It is possible that continual retraining on the system is necessary to ensure that the same observers code the same behaviors in the same way over time.

Interrater Agreement.

Estimates of the extent to which different observers code the same behaviors in the same way indicate satisfactory interrater agreement. Furthermore, these data show quite clearly that reliability of the system is improved with two observations and improved again with a third observation. This is the case whether the multiple observations refer to different observers rating the same lesson or to different observers rating different lessons. Since this is the case, and since there is likely to be more variability in teachers' lessons than in observers' judgments, the importance of multiple observations cannot be underestimated. The most reliable set of scores can be obtained using different observers observing different lessons, but even the same observer observing different lessons considerably improves reliability.

Discriminant Reliability

A study of the use of TADS-MTP with two groups (teachers drawn at random and teachers who had been peer-nominated as superior) indicates that the system can adequately discriminate between the two groups providing that multiple observations are conducted. There is a distinct possibility that with single observations, the most effective teachers cannot be distinguished from one another since there appears to be a "ceiling effect" in that particular score distribution. Seven percent of the observations of teachers drawn at random and about thirteen percent of the observations of teachers nominated as superior resulted in the highest possible scores on the instrument. This means, for example, that if funds are available to make merit awards to only five percent of the applicants, and if ony single observations are conducted, then it is not likely that the teachers who should received awards can be identified. Because scores on the instrument vary from one occasion to the next (as mentioned in the previous section), the distribution of mean scores from multiple observations does not show the same ceiling effect.

Practicality

The most convincing evidence of the practicality of the system is the extent to which it (and the parent TADS) have already been accepted and institutionalized in the Dade County Public Schools. Administrators and subject area specialists have been trained as observers, teachers have been oriented to the system, and over 15,000 observations have been conducted to date.

Score Norms

TADS-MTP does not adequately provide for the comparison of an earned score with the scores of instructors who are in similar instructional situations. It is important to see if there are mean score differences between subgroups of a stratified random sample of teachers. Subgroups of the sample should include the main categories of teaching specialty (elementary, junior high school, high school, major subject areas, etc.). The format and scoring system of TADS-MTP suggest that the more variability that is built into a lesson, the higher the score for that lesson is likely to be. Since teaching behaviors in any given period of time are typically more varied in elementary than in high school classrooms, this could result in typically higher (but not necessarily more meritorious) score for elementary teachers than for high school teachers. This potential bias need to be investigated and, if present in fact, taken into account by weighted scoring. The potential for other subgroup score differences in supported by the results of a survey of Dade County educators which indicate that some subgroups have somewhat different conceptions of what constitutes effectiveness for some of the items.

Behavioral Specificity

(See comments previously made under Behavioral Indicators.)

Factor Unidimensionality

A factor analysis study of TADS-MTP indicates that there are three strong and one marginal but satisfactory factor underlying scores on the system.

Freedom from Bias

The survey of Dade County educators indicates that teachers judge the system to be free of any bias against any person due to race, ethnicity, sex, religious, or any other factors unrelated to meritorious teaching.

Standardized Procedures for Data Collection, Data Analysis, and Evaluation of Performance Measures

Procedures for all aspects of TADS-MTP appear to be well standardized. These include procedures for the conduct of observations, procedures for orientation and training to the system, and procedures for scoring.

Provisions for the Preparation and Approval of Reliable Observers

The history of TADS in Dade County suggests that observers can be trained to a reliable standard on the system, and both the interrater agreement and the fact that so many people have already been trained confirm it. There are established procedures for training both observers and the trainers of observers and these procedures are supported by written and audio-visual materials.

Recommendations

While the review of TADS-MTP against the criteria specified in the Florida state rule reveal several limitations of potential limitations in the system, the team finds that, on the balance, it meets the criteria to a sufficient degree to warrant its use in the Florida Merit Teacher Program for a period of one year. Based on the results of both a reliability study and a study done to assess the ability of the system

to differentiate superior teachers from all teachers, the team strongly recommends that its use in the Merit Teacher Program be conditional on the incorporation of at least two and, ideally, three observations to derive mean scores for merit candidates.

The team further recommends that continuing use of TADS-MTP in the Merit Teacher Program after one year be conditional on the completion of additional developmental work on the system. As indicated in the preceding review, both a normative study and a direct assessment of the stability of observers' ratings are needed. The present evidence based on multiple observations of the teacher on the same occasion is inadequate. In addition, the reliability study that was done used a fairly small sample. It should be repeated with new samples to see if the reliability estimates are stable.

While the reliability and norm studies are important, continuing work on the validity of the system is essential and, because of the nature of the current knowledge base and history of the system, may be a bit more difficult to achieve. The literature review should be updated immediately, the system amended accordingly, and a plan devised for making future adjustments to the system as other research results accrue. This will be extremely difficult to accomplish unless there is a revision in the system to address the need to move away from scoring based on summary inferences and toward the coding of behaviors that have been described with greater specificity. A plan for accomplishing this kind of revision should be addressed immediately. Not only will this improve the chances of ensuring and maintaining the validity of the system, it will also decrease the system's dependence on training procedures, permit incorporation of negative content into some items to control for response set, and ultimately improve confidence in the objectivity of the observations. There is also the distinct possibility that such a revision would minimize the potential for ceiling effects that could preclude differentiation between teachers who achieve scores at the upper end of the score distribution.

The limitations and recommendations concerning TADS-MTP made in this report should, in no way, detract from the significant achievement of the Dade County teachers and administrators who have devised and implemented the TADS system. They are, indeed, to be congratulated for an effort that has had many salutary effects on the teacher evaluation process in the Dade County Public Schools and the process could well be emulated in other school systems wishing to improve their teacher observation and evaluation procedures. The system, as it stands, excels in codifying the conventional wisdom of the profession and starts to link that knowledge base to existing research about effective teaching. If the limitations with regard to its use in the Florida Merit Teacher Program that have been pointed out in this report are successfully addressed, then the system also promises to be an outstanding example of the evolution of educational practice based on research results.

BIBLIOGRAPHY

Evertson, C., Emmer, E., Sanford, J., and Clements, B. (1983) Improving classroom management: An experiment in elementary classrooms. Elementary School Journal, 84, 173-188.

Rosenshine, B. (1983) Teaching functions in instructional programs. Elementary School Journal, 83, 335-353.

Stallings, J., Needels, M., and Stayrook, N. (1979) How to Change the Process of Reading in Secondary Schools, Menlo Park, CA: SRI International.

Appendix M

Comparative Study of FPMS/TADS

Frequencies and Descriptive Statistics

The following appendix contains frequency distributions and major descriptive statistics for key variables produced in this study:

1. a score representing the difference between an individual's ranking on FPMS and TADS,
2. the total score for the TADS instrument,
3. the total score for the FPMS instrument, Master Teacher version,
4. grade levels, and
5. subject areas.

TABLE 63

Frequency Distribution - Difference Between TADS and FPMS

CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT
0.	1	2	2	14.	4	6	47	38.	1	2	79
1.	2	3	5	15.	1	2	48	39.	2	3	82
2.	4	6	11	17.	2	3	52	42.	1	2	83
3.	1	2	12	20.	1	2	53	45.	1	2	85
4.	1	2	14	21.	1	2	55	46.	2	3	88
5.	6	9	23	22.	4	6	61	48.	1	2	89
7.	1	2	24	23.	3	5	65	49.	1	2	91
8.	3	5	29	25.	2	3	68	58.	1	2	92
9.	2	3	32	26.	1	2	70	60.	1	2	94
10.	1	2	33	27.	2	3	73	63.	2	3	97
11.	3	5	38	30.	1	2	74	69.	1	2	98
13.	2	3	41	34.	2	3	77	77.	1	2	100

MEAN	22.394	STD ERR	2.348	MEDIAN	17.000
MODE	5.000	STD DEV	19.075	VARIANCE	363.841
KURTOSIS	0.348	SKEWNESS	1.042	RANGE	77.000
MINIMUM	0.0	MAXIMUM	77.000		

VALID CASES	66	MISSING CASES	0

TABLE 64

Frequency Distribution - TADS Total Scores

CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT
10.	1	2	2	47.	2	3	33	63.	2	3	74
28.	2	3	5	49.	1	2	35	65.	2	3	77
29.	1	2	6	50.	2	3	38	66.	1	2	79
31.	1	2	8	51.	1	2	39	67.	1	2	80
33.	1	2	9	51.	1	2	41	67.	1	2	82
37.	2	3	12	54.	1	2	42	68.	1	2	83
38.	1	2	14	54.	1	2	44	69.	1	2	85
39.	2	3	17	55.	2	3	47	70.	3	5	89
39.	1	2	18	55.	2	3	50	70.	1	2	91
40.	1	2	20	57.	2	3	53	71.	1	2	92
42.	1	2	21	57.	2	3	56	71.	1	2	94
43.	2	3	24	59.	2	3	59	72.	1	2	95
44.	1	2	26	61.	2	3	62	72.	1	2	97
44.	1	2	27	62.	1	2	64	79.	1	2	98
46.	1	2	29	62.	1	2	65	80.	1	2	100
47.	1	2	30	63.	4	6	71				

MEAN	53.864	STD ERR	1.736	MEDIAN	55.250
MODE	62.500	STD DEV	14.100	VARIANCE	198.803
KURTOSIS	0.159	SKEWNESS	-0.584	RANGE	70.000
MINIMUM	10.000	MAXIMUM	80.000		

VALID CASES	66	MISSING CASES	0

TABLE 65

Frequency Distribution - FPMS Total Scores

CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT
51.	1	2	2	59.	3	5	36	67.	3	5	77
52.	1	2	3	59.	2	3	39	68.	1	2	79
53.	1	2	5	60.	1	2	41	69.	2	3	82
53.	1	2	6	60.	1	2	42	69.	1	2	83
54.	3	5	11	61.	3	5	47	70.	2	3	86
54.	2	3	14	62.	3	5	52	70.	2	3	89
55.	1	2	15	62.	3	5	56	71.	3	5	94
55.	1	2	17	63.	2	3	59	72.	1	2	95
56.	2	3	20	63.	1	2	61	72.	1	2	97
56.	1	2	21	64.	2	3	64	73.	1	2	98
57.	3	5	26	65.	2	3	67	75.	1	2	100
58.	2	3	29	65.	2	3	70				
58.	2	3	32	66.	2	3	73				

MEAN	61.879	STD ERR	0.754	MEDIAN	61.583
MODE	53.500	STD DEV	6.127	VARIANCE	37.539
KURTOSIS	-0.965	SKEWNESS	0.158	RANGE	24.000
MINIMUM	50.500	MAXIMUM	74.500		

VALID CASES	66	MISSING CASES	0

TABLE 66

Frequency Distributions for Grade Levels

GRADE	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
.	1	.	.	.
0	13	13	8.667	8.667
1	12	25	8.000	16.667
2	8	33	5.333	22.000
3	11	44	7.333	29.333
4	13	57	8.667	38.000
5	10	67	6.667	44.667
6	3	70	2.000	46.667
7	13	83	8.667	55.333
8	14	97	9.333	64.667
9	17	114	11.333	76.000
10	14	128	9.333	85.333
11	16	144	10.667	96.000
12	6	150	4.000	100.000

TABLE 67

Frequency Distributions for Grade Levels

ELEMENTARY OR POST ELEMENTARY

LEVEL	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
ELEMENTARY (K-6)	67	67	44.667	44.667
POST ELEMENTARY(7-12)	83	150	55.333	100.000

TABLE 68

Frequency Distributions for Subject Areas

SUBJECT	FREQUENCY	CUM FREQ	PERCENT
LANG ARTS/SOC STD	71	71	47.333
MATH OR SCIENCE	40	111	26.667
PE OR ROTC	10	121	6.667
BUSINESS, DCT	7	128	4.667
HOME EC., ART	6	134	4.000
SPECIAL EDUCATION	7	141	4.667
OTHER	9	150	6.000